Objective Assessment Application for Preschool Child Development

by

Guillaume Odendaal

Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Mechatronic) in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof. D. Van Den Heever
Co-supervisor: Prof. P.E. Springer

March 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2020/11/19 .....

Abstract

Objective Assessment Application for Preschool Child Development

G. Odendaal

Department of Mechanical and Mechatronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MEng (Mech)

March 2021

The need for early developmental screening tests is made clear throughout literature. The problem with current developmental screening tests is that they are susceptible to bias and are time- and resource-intensive. Current tests have the administrator convey instructions to a child and then note the proficiency w ith w hich t he c hild h as c ompleted t he t ask, t hus l eaving room for subjectivity and bias within the tester's decision. The lack of trained personnel in rural areas - such as in South Africa - only adds to the sparsity of assessment tools being used. Current tablet assessment applications address these problems but confine t hemselves t o o ne o r t wo m etrics p er construct measured.

Fine-motor and language tests were gathered from literature and standardised tests and implemented on a tablet application. These tests were filtered according to implementability and counsel of medical professionals in the field of early child development. The tablet application was built with modularity in mind to ease the process of adaptation for cultural and age-appropriate conversions. An accompanying assessment pipeline was constructed to automatically process the data from the tablet assessment into interpretable results.

Uittreksel

Objektiewe Assesseerings Toepassing vir Voorskoolse Kinderontwikkeling

("Objective Assessment Application for Preschool Child Development")

G. Odendaal

Departement Meganiese en Megatroniese Ingenieurswese,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MIng (Meg)

Maart 2021

Die nood vir vroeë ontwikkelingstoetse word duidelik gemaak regdeur literatuur. Die probleem met huidige ontwikkelingstoetse is dat dit vatbaar is vir partydigheid en is tyd en hulpbron intensief. Huidige toetse laat die assessor instruksies oordra en besluit dan met watter vaardigheid the kind die opdrag uitvoer, dus is daar ruimte gelaat vir subjektiwiteit en vooroordeel binne in die besluit van die assessor. Die gebrek aan opgeleide personeel in landelike gebiede - soos in Suid Afrika - dra net by tot die ylheid van assesseringsinstrumente wat gebruik word. Huidige tablet assesseerings toepassings adresseer hierdie probleme, maar beperk hulself tot een of twee maatstawwe per konstruk wat gemeet word.

Fynmotoriese en taal toetse is uit literatuur en gestandaardiseerde toetse versamel en op 'n tablettoepassing geïmplementeer. Hierdie toetse is gefiltreer volgens implementeerbaarheid en advies van mediese beroepslui op die gebied van vroeë kinderontwikkeling. Die tablet toepassing is gebou met die oog op modulariteit om die proses van aanpassing vir kulturele en ouderdomsgepaste omskakelings te vergemaklik. 'n Bygaande assesseeringspyplyn is opgestel om die data van die tablet assesseering outomaties in interpreteerbare resultate te verwerk.

Acknowledgements

I would like to express my sincere gratitude to the following people: My mother and father (Nina and Willem Odendaal) for their support during my masters, Professor Dawie van den Heever for all the help and guidance he has given me, and Professor Priscilla Springer for helping with all things medical. I would also like to thank Dr Adri van der Walt for her work and advice with regards to developmental testing, as well as Amy Rode, Romene De Beer, and Cornellia van der Merwe.

Dedications

Hierdie tesis word opgedra aan my familie wat my altyd ondersteun.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Numerous factors contribute to children not developing correctly. Poverty (Duncan and Brooks-Gunn, 2000), malnutrition (Sudfeld *et al.*, 2015), their environment (Evans, 2006), and many more factors can result in a child not developing correctly. On top of this, neurodivergence and disabilities can further stunt development and prevent a child from reaching their potential. Intervention programs have been shown to help children (Jin *et al.*, 2007) if applied effectively. In order for intervention programs to be administered effectively, awareness of problem areas is required. Developmental assessments help by identifying developmental deficiencies and neurodivergence. Classical developmental assessments are done by hand using pen, paper, and a trained medical professional. These assessments can be time- and resource-intensive as well as costly. Additionally, a scarcity of people able to administer the test adds to the difficulty of having widespread use of these tests. Most of the assessment tests have the test administrator instruct the participant to do a task and then scores the participant on how well they did it. This assessment method leaves room for subjectivity and bias to influence the results. Tablet-based assessments are starting to address the subjectivity, but are also not fully utilising the data gathering capabilities of these devices.

1.2 Project Outline

Detailed here within is the process of creating a tablet application able to administer a test and process results from that test. The tablet test in question is a developmental screening test aimed at assessing preschool children's abilities with regards to fine-motor and language. The test has eighteen items, eight for fine-motor and ten for language. All test items were sourced from literature and various other developmental assessment tests, filtered according to implementability and consultation with medical professionals in the field of

child development. The test items are implemented to record as much data about the participant's interactions with the test as possible. Moreover, the test items are constructed with modularity in mind. The processing pipeline receives the data from the tablet assessment test and processes the data into various metrics for interpretation.

1.3 Hypothesis

The fine-motor and language skills of a preschool child can be objectively measured and automatically scored utilising a computerised tablet test and accompanying analysis program.

1.4 Aims

To develop an assessment application for the use in screening preschool children with regards to their fine-motor and language abilities.

1.5 Objectives

The objectives are listed as follows:

1. Gather appropriate test items from literature and pre-existing tests
2. Filter test items based on implementability and suggestions from medical professionals
3. Develop application on tablet implementing all test items
4. Create data analysis pipeline to analyse the data from each of the test items
5. Verify that the application and processing pipeline works as intended

The tablet application will be a series of tasks (known as test items) that the child has to complete. The participant will sequentially complete each test item, and once finished, the test will conclude. The data gathered will then be given to the data processing platform, and interpretable result yielded.

Chapter 2

Literature Overview

2.1 Introduction

In order to develop correctly, a child needs the correct stimulation and care throughout childhood. Correct development is how the child will achieve their full potential, which differs for every person. Children even from birth should be continuously stimulated at their level (Agyei *et al.*, 2016), for example nudging them to crawl and turn over from their backs to their stomach at an early age, or reading to them and mouthing words to encourage their first words. This stimulation happens typically at home or child care facilities such as crèches or nurseries. For any child, the preschool years are the most important, because of increased brain plasticity levels which make it easier for the brain to adapt and change (Chugani, 1998). Children who do not receive the right stimulation and care are at risk of not developing to their full potential. The World Health Organization estimated in 2016 that 43% of children in low- or middle-income countries (250 million) were unable to reach their full developmental potential because of a lack of correct stimulation and care (WHO, 2018). The presence of neurodivergence in some children increases this risk. Neurodivergence is defined as having a brain that functions in ways that diverge significantly from the dominant societal standards of "normal", for example, disorders such as Autistic Spectrum Disorder, Attention Deficit Hyperactive Disorder (ADHD), and dyslexia are considered to be neurodivergent. These disorders can severely affect the development process as memory, attention, lack of emotional and physical control impedes normal development. Intervention strategies can mitigate the effects of these neurodivergent disorders, as well as lack of correct stimulation and care (Steven Barnett, 1998; Gorey, 2001). In order to be able to curate and apply intervention strategies on children in need effectively, awareness of the problem needs to be obtained. Caregivers or guardians/parents are usually the first to become aware of the problems. If lack of correct stimulation and care, presence of a neurodivergent disorder, or both, is discovered early enough, intervention strategies can

mitigate the effect. In South Africa, factors such as poverty, illiteracy of parents, and scarcity of resources can increase the lack of necessary stimulation and care children receive (Engle and Black, 2008). These factors can also contribute to the delayed discovery of neurodivergent disorders. Therefore, affordable assessment tools need to be created and used to increase the detection and awareness of neurodivergence and developmental delays.

2.2 Models of Cognition and Cognitive Domains

Neurodivergence can be measured in different ways. It can be determined and measured by directly observing the brain using, but not limited to, functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and electroencephalography (EEG). These techniques monitor in real-time and are used to detect lesions and underdeveloped regions of the brain (Brown and Jernigan, 2012). The way developmental assessment batteries measure and identify neurodivergence is by measuring the functionality and effectiveness of functional cognitive domains and subdomains. Although most domains and subdomains are agreed upon, there are some inconsistencies.

According to Harvey (2019), eight cognitive domains can be measured, and each domain itself consists of subdomains down to basic component processes. These domains are sensation, perception, motor skills and construction, attention and concentration, memory, executive functioning, processing speed, and language/verbal skills. Sensation refers to a person's ability to detect a stimulus with one their senses, and perception pertains to the processing of this sensory information. Motor skills and construction encompass large and small movements, planning of these movements, and the ability to copy or draw everyday objects. Attention and concentration refer to a person's ability to focus their attention and sustain it. Memory encompasses all facets related to it, such as short-term memory, phonological memory, and muscle memory. Executive functioning is commonly known as reasoning and problem solving, whereas processing speed is the ability to perform simple/complex tasks that require rapid performance. Finally, language and verbal skills are the ability to receive and produce language and to understand and express using language. The domains mentioned can be divided into more general domains containing general processes or brains specific functional models. The general domains are language, executive functioning, memory and attention, and the more specific ones are motor skills and construction, perception, processing speed, sensation.

Baron *et al.* (2012) in their overview of neuropsychological assessment of preschool children consider only six cognitive domains, intelligence, executive functioning, attention, language, motor skills, and memory. Intelligence was considered to be the single best predictive value by which children's develop-

ment could be measured (Baron and Leonberger, 2012). Intelligence is now seen as just one domain out of many that need to be tested. Intelligence testing related to many of the testing methods used for executive functioning, such as problem-solving and shifting. General knowledge is also considered an essential part of intelligence.

Furthermore, the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) and the team of people that compiled it, identified six cognitive domains of importance when talking about impairment and neurodivergence. These domains are perceptual-motor function (the motor skills as mentioned earlier), language, executive function, learning and memory, complex attention, and social cognition (Sachdev *et al.*, 2014). Similar to the description, as mentioned earlier of all domains, social cognition is added. Social cognition relates to one's ability to read social cues, act socially acceptable, read facial expressions, express empathy, and change behaviour based on feedback given.

Lastly, Sabanathan *et al.* (2015) listed five domains that are necessary to test within the context of developing children: cognitive, language, motor, social and emotional, and adaptive behaviour. The first of the domains relates to cognitive processing and is said to include memory, attention, and executive function - more specifically, cognitive flexibility, goal setting, and information processing. Language relates to both receptive and expressive language, and motor relates to both fine- and gross-motor. Social and emotional closely relates to the DSM-5's social cognition. Finally, adaptive behaviour is defined as the collection of conceptual, social and practical skills that have been learned by people in order to function in their everyday lives (Sabanathan *et al.*, 2015).

Although much disparity exists in the definition of cognitive domains, there are a few common domains present. These domains are attention, memory, executive functioning, language, and motor skills.

Attention and concentration consist of selective attention and sustained attention, the former being the ability to ignoring non-relevant information and focussing on important information. The latter refers to how long one's attention on a particular task can last. Attention is a challenging domain to assess in children, specifically younger ages, as their attention-span has not yet fully developed. Attention then becomes an increasingly important aspect of measuring as children who have lesser attention ability may suffer from attention deficit hyperactive disorder (ADHD), or other neurodevelopmental disorders. Attention is listed, mentioned, and reviewed apart from executive functioning for the impact and number of reported neurodevelopmental attentional disorder problems present in the current youth population. Attention is an essential life skill; it is vital to measure the development thereof in preschool children.

Memory, the largest of the cognitive domains according to Harvey (2019), contains many subdomains, namely working memory, explicit memory, procedural memory, semantic memory, and prospective memory. Some subdomains, such as explicit memory, contain sub-processes within them like encoding, storage, and retrieval. Working memory is one's ability to hold information and

manipulate it. Explicit memory deals with the long term storage of information by encoding, maintaining, and retrieving it. Procedural memory relates to the memory of actions and skills, such as muscle memory. Semantic memory is the long term storage of verbal information, and prospective memory is the ability to remember to perform tasks in the future.

Executive functioning is the ability to execute a cognitive set, mental flexibility, and inhibition. A cognitive set is a set of rules to follow or execute when given a task, such as sorting cards into different piles according to their colour. It can be seen as the problem solving and reasoning domain that uses many of the other domains and subdomains to be able to complete a task. The ability to perform a task according to a particular set of rules and then follow a different set of rules measures a person's shifting ability, for example sorting cards according to colour and then according to number. Inhibition is generally seen as one's ability to not act on one stimulus and then promptly act on another. An example would be the tap-knock where if the person administering the test taps, the participant must knock, and if the person knocks then the participant must tap. Goal setting, cognitive flexibility, attention control, and information processing are all considered to be a part of executive functioning (Anderson, 2002).

Language is one's receptive and productive/expressive abilities, the ability to understand language, convey meaning, and follow instructions. Language skill can be measured in a variety of ways such as, but not limited to, measures of fluency (naming as many animals as possible), object naming, and responding to instructions (Harvey, 2019).

Motor skills encompass fine-motor and gross-motor, and more simple constructs such as manual dexterity and motor speed. Construction is one's ability to construct or reconstruct and object (by drawing for example) from either memory or a presented picture.

As the assessment application only measures motor function - specifically fine-motor - and language skills of developing children, a more in-depth explanation thereof will follow.

There is no single domain that needs to be assessed above others; all domains need to be assessed together, if possible. Two domains were selected for the tablet assessment test, fine-motor (a subdomain of motor skill) and language. These domains were chosen for their development period, and affect on long term development and well being. Motor skills are the first to develop and mature in a child (Casey *et al.*, 2005). The early development of motor skills is crucial as assessment should aim to be administered as early as possible, but still be meaningful as assessing a domain that has not developed yet, might not yield usable results. Furthermore, deficiencies in motor skills can have a severe impact on other domains, as well as the quality of life (which will be discussed in section 2.3.1). Language is also one of the earliest domains to develop, developing before other higher cognitive functions such as executive functioning (Richmond *et al.*, 2016). Again, this is important as

assessment should aim to measure domains as early as possible. Language is also seen as one of the most prevalent problem areas in the context of South Africa children(Laughton *et al.*, 2010; Van Der Walt, 2019).

2.3 Motor Function and Language Skills

2.3.1 Motor Function

Motor function consists of two sub-categories, namely fine-motor and gross-motor. Gross-motor is seen as larger movements and one's ability to move the body through the environment. Examples of gross-motor skills are: balance, walking, catching/throwing a ball. Fine-motor correlates more to smaller movements made by the hand such as typing, writing, and tracing one's finger over a line. It is the precision movement of any limb (hands, feet, wrists, fingers). The first areas to mature in one's life are those responsible for motor and sensory processes (Casey *et al.*, 2005).

Development of motor skills follows three basic rules according to Newton and Joyce (2012), cephalocaudal, proximodistal, and gross to specific. Cephalocaudal refers to the development of motor skills from the head to toes, for example, head movement before hand or feet movement. Proximodistal states that limbs closer to the body develop before those further away, such as upper arm control developing before finger control. Finally, gross to specific specifies that larger gross-motor movements develop before fine-motor movement. Within the brain structure, motor function mostly resides in the motor cortex, supplemental motor area, and the premotor cortex.

Impairment in motor function has been seen to accompany a variety of disorders. Dewey *et al.* (2007) set out to examine gestural and motor difficulties in children with autism spectrum disorder (ASD), attention deficit hyperactive disorder (ADHD), and developmental coordination disorder (DCD). Gestural performance refers to the skill with which gestures are performed, such as waving goodbye. Their findings were that children with ASD, DCD, and both DCD with ADHD showed impairment in their motor ability, but only ASD children showed impairment in gestural performance. The researchers noted that the impairment in gestural performance from ASD children might in part be attributed to deficits in language processes as they could have not understood the instructions.

Pitcher *et al.* (2003) more closely examined ADHD and three subtypes thereof to determine accompanying motor difficulties. The three subtypes of ADHD considered were predominantly inattentive (ADHD-PI), hyperactive/impulsive (ADHD-HI), or combined (ADHD-C). They concluded that children with ADHD had significantly lower gross-motor ability and that a high percentage of the children showed motor difficulties consistent with that of DCD children. Of the three subtypes, the ADHD-PI and ADHD-C groups

had a higher impairment score than the ADHD-HI group, but the ADHD-HI group still had a higher score than the control group. Concerning fine-motor, the ADHD-PI and AHDH-C groups had significantly lower fine-motor ability compared to the ADHD-HI and control group. The lower fine-motor ability, they commented, was most probably because of the need for attention in fine-motor tasks.

Furthermore, it has been seen to accompany learning disabilities, such as dyslexia. Fawcett and Nicolson (1995) tested motor skills in children with dyslexia of three different age groups. Three age groups were chosen to be able to see whether or not the results are consistent across ages, which might indicate persistent problems. Children with dyslexia performed the tasks given slower than children without dyslexia of the same age. This difference in speed suggests that children with dyslexia also have accompanying and persistent motor deficits.

Lastly, motor impairment - explicitly relating to balance - has been shown to accompany anxiety disorders (Erez *et al.*, 2004). The researchers strived to investigate the prevalence of balance disorders in childhood anxiety. A group of children diagnosed with general or separation anxiety, along with a control group, were tested for clinically relevant vestibular impairment through extensive neurological examination. What they found was that the group diagnosed with general or separation anxiety made more balance mistakes and had a slower performance than the control group on more challenging balance tasks, such as two-leg balancing on an unsteady surface, or one-leg balancing on an unsteady trampoline. The researchers further mentioned that it might be the anxiety that is causing balance dysfunction - as a psychosomatic manifestation - or the balance dysfunction may be causing the anxiety to manifest.

Along with motor impairment accompanying other disorders, motor dysfunction on its own can be classified as a disorder. The American Psychiatric Association defines Developmental Coordination Disorder (DCD) as a marked impairment in the development of motor coordination that significantly interferes with academic competence or daily living skills.

The effect DCD has on children's mathematical skill, reading, and working memory was investigated by Alloway (2007). What the results indicated was that children with DCD have significantly worse visuospatial memory than verbal short-term memory. This difference in memory, according to the researchers, is consistent with previous research linking visuospatial memory to movement planning and control. Worse visuospatial memory, in turn, is linked to worse memory and learning ability. Moreover, motor impairment was found to impact social (Smyth and Anderson, 2000), and emotional (Cairney *et al.*, 2010) functioning as well.

Piek *et al.* (2012) mentions the importance of uncovering motor developmental issues before the commencement of school as motor disorders, or related disorders, may impact the child negatively when they are unable to

complete specific motor-related tasks. The result of this can be scarring on children, leaving them with damaged self-images, and sometimes have them avoid social interactions.

A few longitudinal studies have found a relationship between early motor ability and the performance of certain domains later in life. Earlier acquisition of motor abilities was linked to better adult executive functioning later in life (Murray *et al.*, 2006). Furthermore, the early performance of motor ability was linked to later academic - specifically mathematics - performance (Kurdek and Sinclair, 2001). Lastly, preschool motor ability was able to predict levels of anxiety and depressive symptomatology at school age(Piek *et al.*, 2010).

Assessment of motor ability is necessary when developmental assessment is being undertaken. Both gross and fine-motor assessments are necessary. However, this is not possible with a tablet application. Most test items testing gross-motor would be challenging to transfer to a tablet assessment platform but is still possible. Therefore, the scope of the assessment is shifted to only testing fine-motor ability.

2.3.2 Language

Language is an integral part of how humans interact with the world. We read, speak, and think in a language. Through a particular language, a child learns everything they can about the world; they express themselves and interact with other people. Language is vital in our modern society.

The acquisition of language starts shortly after birth when infants can discriminate between different sound contrasts (McMurray and Aslin, 2005). From birth up until the first word is known as the prelinguistic period, whereby the infant will start to make speech sounds, babbling, and longer sequences of sounds trying to mimic adult speech (Saaristo-Helin *et al.*, 2011). The next developmental phase relates to gestures where simple gestures are used to indicate wants and interactions (Behne *et al.*, 2012). Basic language comprehension marks the final segment of the prelinguistic period with the infant starting to respond to his/her name and associating words with object (Tincoff and Jusczyk, 1999). After the first word is spoken, language acquisition accelerates with the infant acquiring on average, ten words per month. The acquisition of ten words per month continues until the child's vocabulary has reached the size of about 50 words, whereby word acquisition rate increases to over 30 new words per month (Goldfield and Reznick, 1990). Two-word speech indicates the basic grammatical knowledge developing (Schipke and Kauschke, 2011) which in turn becomes three- and four-word utterances. At this stage, auxiliary verbs follow shortly after, but questions and negative sentences follow later (Tyack and Ingram, 1977). Finally, the language development phase is complete when children reach the end of their preschool years (Hoff, 2009).

The complex nature of language can be described as a system comprising of many dimensions, namely phonology (the sound system), the lexicon (the

vocabulary), semantics (meaning), grammar (structure), pragmatics (communicative functions and conventions for language use), and discourse (the integration of utterances into longer stretches of conversation or narrative) (Conti-Ramsden and Durkin, 2012). All these dimensions together make up the systematic model of language, and it is to be noted that assessing only one dimension of a child's language can give a false perception of the child's language profile. Therefore a variety of dimensions in the systematic language model need to be assessed.

There are two subdomains in the context of a person's language, namely expressive language and receptive language. Expressive language is the ability of a person to communicate their needs and wants. Receptive language is seen as the comprehension of language, how well a person understands the message conveyed to them, how well they understand the message (attention, ability to hear), and how well they process said message (following directions, and understanding questions).

The primary detector for language-specific disorders is a caregiver, or parent (the guardian of the child). Signs of deficiency might go unnoticed for a long time, if at all before the guardian realises something might be wrong. The variability in language acquisition makes it challenging to create a robust detection tool for language-specific disorders and deficiencies, and even more so for a diagnostic tool. This variability also plays a role in delaying the guardian reaching out to practitioners able to assess the child. After the preschool period, a child then goes to a school where there are teachers who are trained to identify and notify the guardian(s)/professionals of possible deficiencies. Deficiencies may still go undetected as some classrooms are full and busy, and the child may go unnoticed. With each delay in detection that there is a deficiency, the problem grows and may have adverse effects on the child, such as reading difficulty into adolescent years (St. Clair *et al.*, 2010) difficulty in school (Conti-Ramsden *et al.*, 2009), and worse social bonds and relationships (Durkin and Conti-Ramsden, 2007).

There is a need for early identification as children's language growth fluidity is higher at a younger age (Bishop and Edmundson, 1987). Therefore, the need for assessment is to inform early intervention programmes (if necessary), provide needed help to cope with or mitigate the effects of a deficiency/disorder or to create a developmental profile for a more thorough inspection.

Law *et al.* (1998) stated that language tests might struggle because of the varied nature of language acquisition in preschool children, the lack of early and robust language deficit predictors, or the lack of strong identifiable generic or neurobiological markers of language impairment.

Language ability, or the lack thereof, has been linked to various developmental and educational outcomes. More apparent would be the linkage between phonological short term memory, language and literacy (Conti-Ramsden and Durkin, 2007). Moreover, language deficiencies can also affect social behaviour and quality of friendship (Durkin and Conti-Ramsden, 2007). Lastly,

emotional difficulties and academic failure have also been linked to language impairment (St Clair *et al.*, 2011; Conti-Ramsden *et al.*, 2009).

There are several ways to assess a preschool child's language ability, such as using standardised assessments and informal/dynamic assessments. Informal assessments might put the child in question in a more comfortable context, thus revealing more natural and everyday language usages (Schaefer *et al.*, 2016), but it is time-consuming and not always comparable between different participants. More formal standardised assessments give norms whereby a child can be compared to their peers, the context might sometimes be different from their everyday life to the extent where the test's results become unreliable (Conti-Ramsden and Durkin, 2012). Furthermore, there are other considerations when measuring a child's language skill. For one, monolingual and bilingual differences have to be taken into account (Crutchley *et al.*, 1997). Another would be that expressive and receptive language skills need to be separately assessed (Bishop and Adams, 1990; Paul, 1996). Finally, it is insufficient to measure only one of the dimensions of language previously mentioned (Thal and Katich, 1996). Therefore, there is a need for an assessment instrument to measure the multi-dimensional nature of language.

2.4 Classical Development Assessment

Cognitive domains are, in nature, difficult to assess as they are not physical attributes easily discernible from one another such as weight or height (Conti-Ramsden and Durkin, 2012). Classical developmental assessments refer to assessments done utilising pen, paper, and observation. These assessments can include diagnostic tests, where a diagnosis is made after the results have been analysed, or screening tests, which cannot be used to make a diagnostic prediction of neurodivergence but can give good indicators of whether or not further assessment is needed. Diagnostic tests are very time consuming but give very in-depth assessment results, whereas screening tests are quick to administer but cannot be used for diagnosis. The general test structure consists of a series of actions that need to be done or instructions that need to be followed, named test items. The participant must perform actions such as hop on one leg, name the object being pointed at, or sort cards into a pile according to their colour. Before being used as assessment tools for diagnosis, these developmental assessment tests need to be standardised and norm-referenced. Norm-referenced means that the tests have been administered to a large enough group of participants that participants can be ranked and compared to one another. Norm-referencing is necessary to determine when test results are abnormal, and further investigation is needed. Tests typically have a guideline or manual that describes the exact procedure of administration. These guidelines also describe what to look for, how to score tests, and interpret results. While the test is being administered, the admin-

istrator writes descriptions of how the participant is performing each action, notes any possible abnormalities, and transcribes what is said (for language-based assessments). An example of abnormalities that might get noticed is the inability to pronounce the letter **R**. Some tests give scorecards with rating scales from one to three or one to five to help the administrator assess the participant better (the administrator would observe the participant performing the action and rate them on the scale given), or binary rating scales which indicate whether the participant was able to complete the action or not. In order to administer these developmental assessment tests and accurately score the participant, one needs to be a medical professional. Some tests, such as the Griffiths Mental Developmental Scales, require further training in order to administer them effectively. Developmental tests assess cognitive domains separate from one another, with some tests only testing one cognitive domain. Separate scores are calculated, and separate scoring cards are given if more than one cognitive domain is assessed, or even for cognitive subdomains (for example, a score would be calculated for fine-motor skill and gross-motor skill separately, and then combined to form a motor skill score). Furthermore, each domain's assessment results cannot be viewed and interpreted independent of other domains as most domains work together to complete specific tasks. Therefore, it is sometimes necessary to have multiple assessments, or one big assessment battery, performed, and the resulting score analysed.

There are many language assessments, but certain factors need to be taken into account when selecting an assessment test. Concerning language, Conti-Ramsden and Durkin (2012) compiled a list of norm-referenced assessment along with age range, and the size and location of the normative data. Assessments can either assess the general concept of language or specific dimensions thereof. British Picture Vocabulary Scale 3^{rd} edition (BPVSIII) (Dunn and Dunn, 2009) presents the participant with four options and a word verbally spoken. The participant must select the option that best describes the stimulus. Clinical Evaluation of Language Fundamentals (CELF) (Semel *et al.*, 2006) measures numerous language skills such as sentence structure, word structure, expressive vocabulary, concepts and following directions, recalling sentences, basic concepts, and word classes. These concepts are measured in a variety of test items where the participant must repeat sentences, words, name objects presented to them, and perform simple tasks when instructed. The Early Repetition Battery (ERB) (Seeff-Gabriel *et al.*, 2008) is a language assessment test whereby the participant must repeat words, sentences, and non-word sounds back after hearing them. The Expressive One-Word Picture Vocabulary test (EOWPVT) (Martin and Brownell, 2010*a*) requires the participant to name the object placed in front of them. In contrast, Receptive One-Word Picture Vocabulary test (ROWPVT) (Martin and Brownell, 2010*b*) requires the participant to select an option that best describes the stimulus word spoken, similar to BPVSIII. The Expressive Vocabulary Test (Williams, 2007) requires the participant to describe an image/scenario presented. Similar

to others, Peabody Picture Vocabulary Test-IV (PPVT-IV) (Dunn and Dunn, 2007) presents the participant with a stimulus image, and the participant is instructed to describe the image.

Likewise, regarding motor skills, there are numerous assessment batteries and tests. Movement Assessment Battery for Children (MABC-2) (Henderson *et al.*, 2007) is the second and revised version of the first. It estimates both fine- and gross-motor ability by measuring aiming and catching, manual dexterity, and static and dynamic balance. The scoring is done by having the participant throw/catch a ball, balance on one leg, thread beads onto a string, post coins into a mail slot and having an administrator score how well the task was completed. Bruininks-Oseretsky Test of Motor Proficiency 2nd edition (BOT-2) (Bruininks and Bruininks, 2005) is the revised version of the BOTMP (Bruininks-Oseretsky Test of Motor Proficiency by Bruininks *et al.* (1978)) with the aim of more reliable assessment for four to five-year-old children. It also measures both fine- and gross-motor by having eight subtests, namely strength, upper limb coordination, running speed and agility, bilateral coordination, manual dexterity, fine-motor integration, and fine-motor precision. The measurement is done by having the participant manipulate objects (picking up coins and putting them in a bottle, picking up and placing cards), place pegs in a pegboard, drawing objects such as triangles, colouring in circles, and tracing lines with a pencil. At the same time, the administrator observes and grades their performance. Peabody Developmental Motor Scales 2^{nd} edition (PDMS-2) (Folio and Fewell, 2000) is suitable for infants up to children before they attend school. Again, the assessment test instructs the participant to perform tasks and then scores how well the participant completed the task. Measuring both fine- and gross-motor but having different scoring for each of the subdomains, it is still said to be reliable. In contrast to the BOT-2 test, it does not reliably identify minor motor problems (Slater *et al.*, 2010). McCarron Assessment of Neuromuscular Development (MAND) developed by McCarron (1997) aims to measure motor function (both fine- and gross-motor) and contains test items such as threading beads on a rod, placing beads in a box, finger tapping, jumping, and heel-to-toe walking. Once the participant is given the instructions to complete the task, the administrator scores how well the participant completes the task.

Some test batteries are not restricted to one functional domain and measure a combination thereof. The need for early detection and intervention has prompted Aoki *et al.* (2018) to create the Neuromotor 5-minute Exam (N5E) and a different version named the Neuromotor 5-minute 2-year-old version (N5E2). The initial goal was to give medical professionals a short yet effective screening tool. The test items were selected based on being able to indicate neurological abnormalities, can be administered without specialised training, and scoring can be done regardless of the examiner's expertise or background. It measured perception, cognition, language, physical characteristics, tone abnormality, and motor problems.

DDST (Frankenburg *et al.*, 1992) in its entirety measures four categories: Personal-social, fine-motor, language, and gross-motor. It measures these categories by having participants complete test items and rating the participant's test behaviour concerning compliance, interest in surroundings, fearfulness, and attention span. With a focus on the language section, the test contains test items such as defining the composition of materials, defining the meaning of words, giving opposites to words, verbally recognising colours, comprehending prepositions, and following directions. The fine-motor assessment section contains picking the longer of three lines, drawing a stick figure, draw a square/cross/circle, and ability to build a tower using 2/4/8 blocks.

DDST has been adapted several times to accommodate the cultural differences between the West (the U.K. and the U.S.) and others. Examples of countries with such adaptations are Uganda (Nampijja *et al.*, 2010), Kenya, Malawi, and Iran. Nampijja *et al.* (2010) undertook a process to create a developmental screening test which would work in their cultural context. They tested five categories of development: attention, executive function, general cognitive ability, language, and motor ability (fine- and gross-motor). They adapted some tests out of standardised child developmental test batteries such as NEPSY and British Ability Scales.

The Griffiths Mental Developmental Scales (GMDS) contains sub-scales for language and fine-motor as well. The language sub-scale measures both receptive and expressive language through a series of tasks, namely defining an object by use, describing a picture, repeating a sentence given, pointing to and naming objects in a picture, naming opposites, and identifying the composition of objects. The fine-motor sub-scale contains tasks that involve drawing objects, copying objects from an image or memory, and the ability to use scissors.

South African psychologists recommend the Griffiths Mental Developmental Scales for child developmental assessment because of its use in South Africa. To be able to perform the Griffiths assessment on a child, one would need to undergo training. Only paediatricians, psychologists, or allied health professionals who are a part of a child developmental team, actively involved in research or monitoring, or supervised by an experienced Griffiths user can apply for training programs.

The Griffiths assessment, currently in its 3^{rd} revision, has been used in research with regards to South African populations. The Griffiths assessment has been used within South Africa across various cultures (Amod *et al.*, 2007), and for longitudinal studies (Laughton *et al.*, 2010).

These tests are costly to administer, both because of the price of the test, and the time it takes to administer, as medical professionals are required to perform the assessments. Furthermore, some assessments require someone to undergo training before being allowed to acquire and administer the test. This prerequisite training decreases the availability of people able to administer the test and further increases the cost. Another problem is the subjectivity

in assessments. Many of the test item measures are assessed subjectively, whereby the administrator of the test has to score how well the participant is performing the task. The participant would be instructed to perform a specific task such as threading beads on a string, using scissors to cut a piece of paper, using building blocks to build structures, drawing objects on paper, describing a picture, pronouncing a word, or giving opposites. An administrator would then assess how well the participant performed the action and rate it on a scale or merely indicate that it was completed successfully. This manner of scoring leaves room for subjectivity and bias that can affect the results, which can lead to misdiagnoses or incorrect recommendations derived from the results.

2.5 Computerised Development Assessment

As previously mentioned, early developmental assessment of children is essential - as both language and motor development form integral parts of our lives. Deficiencies in either of these areas can lead to problems now and later in life. The current assessment methods require a trained professional to administer an assessment test and record the data manually, usually with a paper and pen. The need for a medical professional makes the process costly and time-consuming, and also inaccessible to low-to-middle-income countries (Pitchford and Outhwaite, 2016). The inaccessibility is compounded when speech and language are being assessed, as it becomes challenging to identify language deficits when the professional administering the test is not as proficient in the participant's home language (Schaefer *et al.*, 2016). Although tablet assessments and other computerised developmental assessments are a step in the right direction to mitigate the subjectivity found in classical assessments, there are still some subjective measures present.

Tablet technology could aid developmental assessment in a wide array of scenarios as it is lightweight and compact (Kucirkova, 2014). Furthermore, young children (2-3 years) can successfully interact with this technology (Nacher *et al.*, 2015) as tablets are familiar since they have been introduced across the world (Chiong and Shuler, 2010; Geist, 2012).

Before touch screen tablet tests can be used as assessment or screening tools, they need to be assessed themselves. The test, and the items it contains, need to be correlated with the results of well-known and widely used normative assessment tests. Furthermore, each test item must be guarded against bias and tested for its validity in what it is meant to measure. Lastly, the test items need to be culturally appropriate for the context within which the participants live (Pitchford and Outhwaite, 2016) and the language(s) the participants speak (Schaefer *et al.*, 2016).

There are currently only a handful of tablet developmental assessment applications. Pitchford and Outhwaite (2016) created a touch screen tablet tool for use in cross-cultural motor and core-cognitive skills assessment. The con-

structs the tablet assessment tool assesses are manual processing speed, manual coordination, short-term memory, visual attention, working memory, and spatial intelligence. These constructs are tested by measuring how long the participant takes to complete the task (manual processing speed and manual coordination) and whether or not the participant was able to successfully tap the correct dots on the screen (visual attention, working memory, and spatial intelligence). The metrics are the time to completion and whether or not the participant has completed the task. These metrics assess the constructs, albeit not very in-depth. Along with the newly created touch screen assessment tool, Pitchford and Outhwaite (2016) used two additional standardised measures with which to compare the results. Block Design and Symbol Search acquired from WPPSI-III were chosen as there are similarities between the touch screen assessment and the standardised one. Furthermore, the Symbol Search test is used to measure cognitive processing speed, which is strongly correlated with working memory, which is, in turn, again correlated with short-term memory. The reliability and validity assessment of the touch screen tool was derived from a series of correlations. Test-Retest reliability was calculated by giving a particular group the tablet test twice, with eight weeks of separation. It resulted in correlations of low to moderate strength ($r < 0.5$) for the test items.

Schaefer *et al.* (2016) created a tablet assessment tool intending to measure English home language and non-English home language children's receptive language skills objectively and reliably. The tool was translated into eight additional languages. The test is structured around one test item: four pictures are presented to the participant, an audible stimulus conveys a word, and the participants have to select the appropriate picture (much like ROWPVT). The words that were used as the audible stimulus were gathered by Kuperman *et al.* (2012) and then filtered based on having easily representable verbs/nouns and not having cultural bias or ambiguity when being translated, having only a single translation, and being culturally relevant. Of the four pictures, one was the correct answer; one was a categorical distractor (e.g. the target would be a book and the distractor would be a newspaper), another a meronymic or functional distractor (the target would be a monkey and the distractor would be a tail, i.e. part of distractor), and the last would be a random distractor. Each of the test items' images was matched with the age of acquisition data to ensure a participant would understand/know all the images. The assessment tool mitigated subjective measures by having the tablet automatically score the option selected by the participant. However, the test only consists of showing an object and recording the option selected. Furthermore, a standardised control test was done with a BPVS and a CELF test in order to validate the newly created tablet test. Low to moderate correlations were found ($0.214 < r < 0.597$) when comparing the created assessment tool to the CELF and BPVS test using monolingual and multilingual groups. Final remarks from Schaefer *et al.* (2016) was to add more measurements to the tablet app (reaction

time, further checking linguistic properties of the test and distractor items) and possibly make the application web-based in order to be able to build an anonymous receptive vocabulary dataset.

Another type of tablet-based test is one developed by Francis-Lyon *et al.* (2017), that is more focussed at easing the assessment procedure from the perspective of the assessor. The problem the researchers addressed was that when professionals administer the developmental assessments (the Kilifi Developmental Inventory in this case), they struggled to convey what needs to be done to the participant and then record the response on pen and paper along with keeping time. Several other problems were mentioned and related to the assessment procedure taking too long, such as children becoming restless, or the test needing to be adapted to avoid skipping entire sections (if a question's answer rendered a section invalid). The measures are still subjectively assessed as the application only facilitates the gathering process. The tablet application is a customisable assessment sheet that can be used to display stimuli (for the children to redraw), keep time (when a timed assessment is necessary), and record audio for later transcription. The customisation is built on giving a text block containing the information about the test to the tablet application, and the tablet generates the test. This customisation is an essential characteristic needed for wide-spread use (across cultures and different ages).

In an attempt to measure self-regulation, executive function, language, and social development objectively, reliably, and with ease of administration, Howard and Melhuish (2017) created the Early Years Toolbox (EYT). More specifically, the measures are visuospatial and phonological working memory, shifting, inhibition, vocabulary, and a parent or guardian report of self-regulation and social behaviour. Visuospatial working memory ability was measured with a "Mr Ant" task whereby a cartoon ant image would have dots displayed on it for a brief period then disappear. The participant would then have to indicate where the dots were placed. A phonological working memory task conveyed an instruction to the participant of what object not to select and then recorded the selected object. With each level the instruction increased in length, adding features for the participant to remember when making a selection. Inhibition was measured with a "Go/No-Go" task where the participant is required to act (touch the screen) on a go signal or refrain from acting in a no-go scene. The signals were 80% go signals and 20% no-go signals to generate a prepotent tendency to act. Shifting was measured with a card sorting task of rabbits and boats. Two sortable categories were made available on-screen, a blue rabbit and a red boat. The participant has to switch between sorting oncoming objects (blue/red rabbit/boat) into specific categories according to the current rule (either by colour or by shape). Lastly, language development was measured using an expressive vocabulary task. The participant is presented with an image portraying a familiar object (familiar in order to negate context bias) and must verbally label the object. This test item still employs a person to listen to what the participant has said and decide whether or not

the participant produced the correct prompt, leaving room for subjectivity.

Similar to other tablet tests (Pitchford and Outhwaite, 2016), Howard and Melhuish (2017) administered other standardised tests to the participants in order to evaluate the convergent validity of the newly created tablet test. List sorting (working memory), inhibition (flanker), and shifting (dimensional change card sorting) from the NIH Toolbox's Cognition Battery were used to correlate their results with a standardised test. Furthermore, the BAS-2 Expressive Vocabulary subtest was also administered. Lastly, internal consistency analysis was conducted on the "Go/No-Go", "Expressive Vocabulary", and questionnaire tasks.

To entirely mitigate subjectivity and reliance on a medical professional to compile the results, Bhavnani *et al.* (2019) sought the use of machine learning algorithms to compile and analyse the results and make predictive assessments. The assessment tool is a game that the participants play, but the analysis is done in the background. Quantitative measures were sought out and conceptually verified by consulting paediatricians, neuroscientists, psychiatrists. The test consisted of nine items measuring various constructs such as inhibition, attention, visual form perception, visual integration, reasoning, memory, manual processing speed, and manual coordination. In the pilot study, Mukherjee *et al.* (2020) administered the newly created tablet test alongside the Bayley's Scale of Infant and Toddler Development version 3 (BSID-III). Data, which was carefully defined using a team of professionals, was used alongside a variety of machine learning algorithms to be able to predict the participant's BSID-III score, acquiring a correlation of $r = 0.67$.

The tablet assessments mentioned mostly counteract the subjectivity problem encountered in classical developmental assessments, but lack in the depth of assessment and data recording available on a tablet device. Classical developmental assessments are constrained to assessing and one or two measures per test item as a person has to observe and write down the observations. Assessment tests on tablets are not constrained by the same limitations and can observe and gather data using numerous parallel processes. The tablet assessments mentioned above also confine themselves to these constraints and only measure one or two metrics per test item.

2.6 Summary

Developmental assessment is of critical importance in order to detect neurodivergence and a lack of sufficient development. It is especially crucial in preschool years where a child's brain can adapt and overcome certain deficiencies, with enough help in the form of intervention plans. However, in order to effectively administer intervention plans, awareness of neurodivergence and developmental delays must be acquired using these assessments.

Assessing all cognitive domains is the ideal case, but this is not always possible. Motor skills is the domain that first develops in a child and can have an impact on other cognitive domains later in life if not correctly developed. The language domain develops before other higher cognitive functions and is seen as one of the most prevalent problems facing South African children. Therefore, the tablet application in question focuses on fine-motor and language domains. These domains are of the first to develop and have a significant impact on cognitive and other domains later in life. Fine-motor, rather than motor as a whole, is explicitly assessed as gross-motor assessments are not possible to transfer to the current tablet application platform.

Classical developmental assessments are done by pen, paper, and a medical professional observing the participant. Instructions are given, and the medical professional scores how well the participant performed the task. These assessments are susceptible to subjectivity and bias, which can influence the results acquired. Furthermore, classical developmental assessments are expensive and time- and resource-intensive to administer.

Tablet assessments are a step in the right direction with regards to limiting the subjectivity of classical assessments, but some subjective measures remain. These assessments only measure one or two metrics per test item, not taking advantage of the parallel processing capabilities of tablet devices.

Chapter 3

Methodology and Implementation

3.1 Introduction

The tablet application is a series of tasks (which is referred to as test items) that the child has to complete. Each test item will be done sequentially, and after all test items are done, the test will conclude, and the data gathered will be given to the data processing platform.

There are a total of eighteen test items present in this tablet assessment test, each selected from literature or gold standard tests and adapted to be viable on a tablet medium. The application was built with modularity in mind to ease the process of adaptations for different contexts. Every image shown, every word pronounced can be changed to suit the researcher administering the test by changing entries in the resource and scenario database, which is explained in a later section. Scenarios define how each test item should behave. It is a set of values that tell the test item what to display, how to display it, and what to do next. Each test item can have a variable amount of scenarios which can be seen as sub-tests. Resource items are the objects the tablet displays, such as a picture of a tree, or a sentence being read aloud to the participant. An analogy to help explain the test item, scenario, and resource item relationship is as follows: Think of a screenplay that has to be performed by actors on a stage. The stage has to be set up in a specific way with props and objects, creating a specific environment; this is the test item. It is a framework set up to house scenarios and resource items to test a specific construct (such as someone's fine-motor ability). For the play to be able to take place, the actors have to know what to do, when to do it, and for how long it should be done. These metrics are all defined by the scenario of each test item. It is similar to how the script of a play would work. There are multiple scenarios per test item, similar to how there are multiple acts in a play. Each of these acts has a script. Finally, the scenario, or script, specifies certain actors to come to the stage at certain times. These actors are resource items and are displayed to the participant (the audience).

This application is a two-part build, the data-gathering application housing all the test items and the data processing application that processes the data and presents the researcher with a variety of results and scores to interpret. Usually, the data processing is done by the person administering the test, but as previously mentioned, this can be vulnerable to subjectivity and depends on the administrator's education and background. Therefore, the data generated from the test items are automatically processed into interpretable results.

The following sections will start by explaining the technical implementations of the tablet application. Then the origin and reason for the inclusion of each of the test items. Subsequently, each test item will be described and mapped out by explaining how it was implemented and how each can vary. After that, the data being logged from each test item is described. Finally, the data processing segment will follow and explain the six categories of analysis present in this application, how each category manipulates the data, which test items' data is being processed, and what results are shown.

3.2 Data Gathering Tool and Test Items

3.2.1 General Setup and Structure

The data gathering application was built for an Android tablet, using Java 8 and a free interactive development environment (IDE) named Android Studio 4.0.1. The options most notable for the development of such an application are Native or WebApp. Native development, which refers to developing an application in a platform-specific language (Android uses Java/Kotlin and iOS uses C#/Swift), was chosen over WebApp development as the features needed were not available in the WebApp frameworks considered. Furthermore, the cost of development for a native application (which could be done in-house, thus no need for contracting a developer) was less than that of the cost of WebApp development. Similarly, Android was chosen as the native platform over iOS as the developmental cost for an iOS application was more than that of an Android Application, which could be done in-house as well.

The application starts at the Home Screen displaying three options, Start, Settings, and Exit. The test battery can be started using the Start button, or the test battery's test items can be selected and removed in the Settings menu. The settings menu, as seen in figure 3.1, lists all possible test items. Tapping a test item adds it to the test battery list, on the right. Tapping an item on the right will remove it from the current test battery.

Android applications work with Activities. Each Activity can be seen as a single screen (although not always the case) with its own lifecycle, meaning it receives its own inputs, displays something on the screen, and outputs data to the Activity that started it, or to a next activity it starts. Fragments, which can be placed on top of an Activity, represents a portion of the user

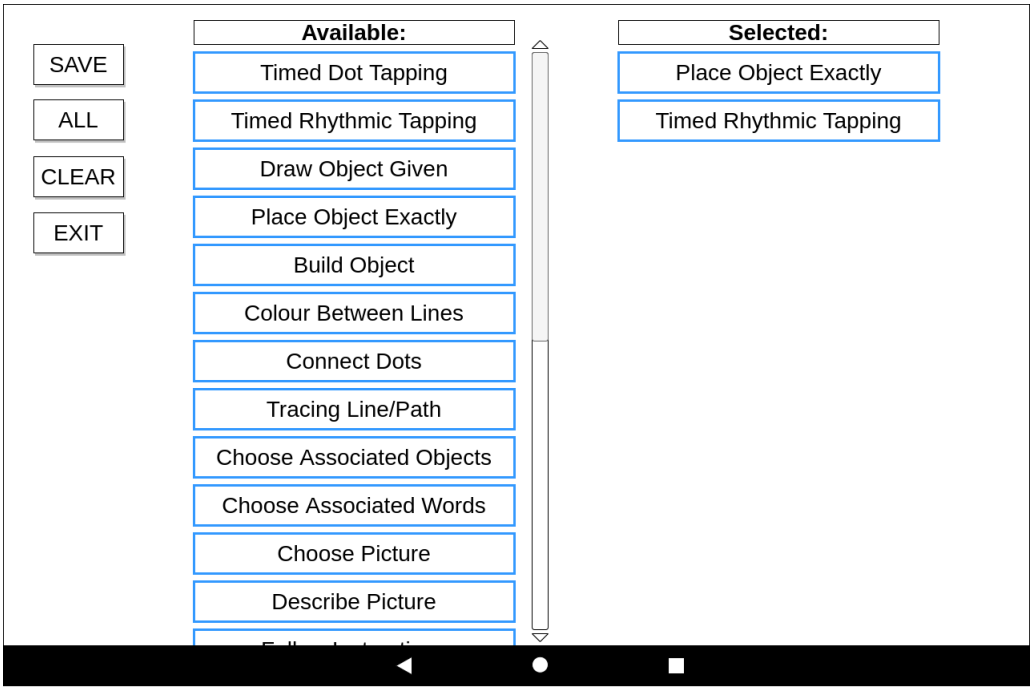


Figure 3.1: Setting screen where items are selected to be used in a test battery.

interface separate from the Activity and also has its own lifecycle. Each of the test items is constructed as fragments as they can easily be interchanged and require fewer resources to run. Figure 3.2 shows the base activity housing the fragment, which in turn houses its components. The Activity that hosts each of the test item fragments has three components, which are used in each of the test items: The BACK button, the DONE button, and the instruction description at the top of the screen. As each test item starts, the Activity receives that test item's instructions and changes the instruction string to display it. The BACK button returns the participant to the home screen, and the DONE button indicates that the participant is finished with the test item.

Each test item is created from a variety of customised components. These components, detailed in section A.2, were created by extending the classes of pre-existing components made available by the Android Studio IDE. By extending these classes, custom behaviour (such as displaying images and sounding words when touched) was implemented to ease the development process. These custom components can receive data from the test item's scenario details, act upon it in a predetermined way, and log everything that happens to the component itself.

The way Android devices assign coordinates to the pixels on the screen starts from the top left-hand corner. When referring to logging the location of a component in the succeeding sections, what is meant is that the top left corner pixel location of the component is being logged. Furthermore, the height and width attributes of the component are also logged.

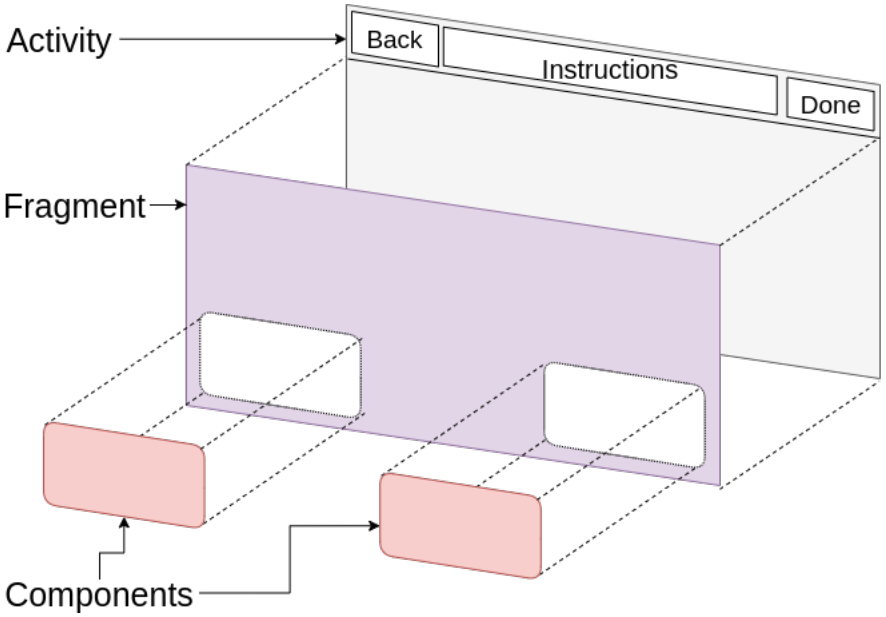


Figure 3.2: Illustration of how fragments, components, and activities fit together in this specific application.

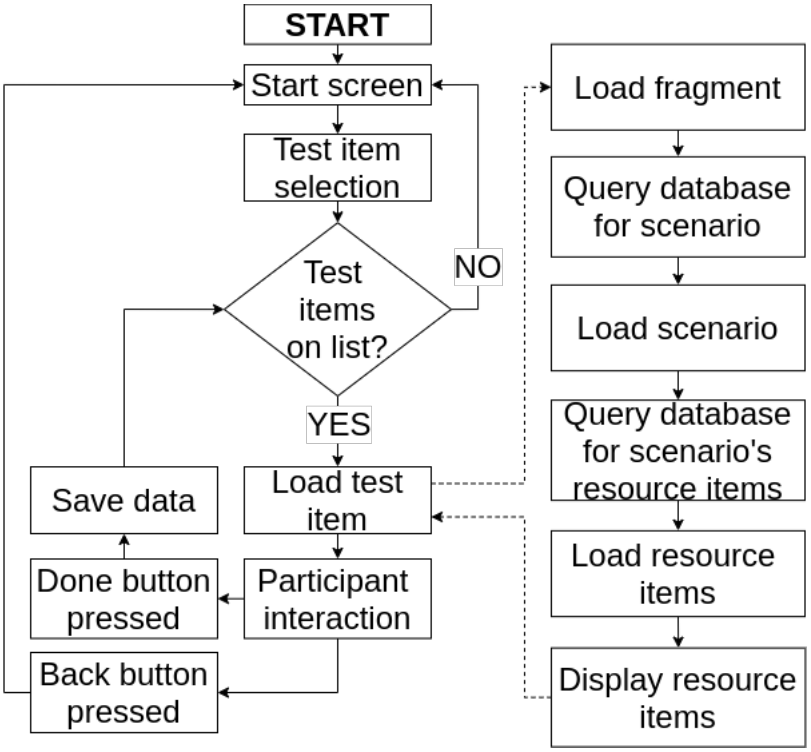


Figure 3.3: General flow of the data gathering application.

Referring to figure 3.3, the application starts at the start screen. The researcher selects the test items to be administered in the settings menu, adding them to the test item list. The test starts by loading the first item on the list, which is done by getting the fragment of the test item ready. The scenario database is then queried for scenarios of the test item in question. Each scenario has a specific set of resource items it uses, and this is acquired by querying the resource items database. The first scenario and its resource items are then displayed to the screen, and the test starts. As the scenario is finished, the next scenario is displayed until there are no more scenarios are left for the test item. The next test item is then loaded until no more test items. Finally, the program reverts to the starting screen.

All data logged by the data gathering tool is stored in a JSON file with a specific format. Although the content may differ from test item to test item, a general structure is upheld. The structure is seen in figure 3.4. The JSON file (denoted as list 1) contains the start date and time of the test, a unique identifying ID and an array of test items that were performed. Each of these test items (denoted as list 2) contains identifying information along with all the scenarios that took place. Every scenario (denoted as list 3) again contains identifying information (see scenario structure) and an array of events that happened while the scenario was active. These events (denoted as list 4) contain identifying information as well as the data that needs to be processed later on. An event is a generic piece of data logged from any component. When a component is displayed, it logs its location, along with its height and width, which is stored in the active scenario's events array. When a component is touched, it logs information about the touch, which again is put into the active scenario's events array.

All logging within the application is done by using broadcasts. Broadcasting in the context of Android applications is the process of sending a public message to every active component in the application, and is synonymous to a person using a megaphone to yell a message to a group of people. Each component, just like each person, would *hear* the message, but will not react unless they were specifically listening for the message. Components can be programmed to listen for particular messages by checking the broadcasted message to see if they were the recipient. To continue with the megaphone analogy, the person yelling through the megaphone would specifically mention the names of people that have to respond to what is being yelled. Each person checks to see if the message contained their name, and reacts in a predetermined way if it did, or ignores the message if it did not. In order to log everything that happens, each of the components can broadcast a message in the application. A logging service in the background listens for specific keywords that identify the type of message and log it in the events array of the current scenario. Each broadcast contains a unique ID to be able to differentiate between components of the same type from the same test item. There are numerous broadcasts created to log the data in the application and can be

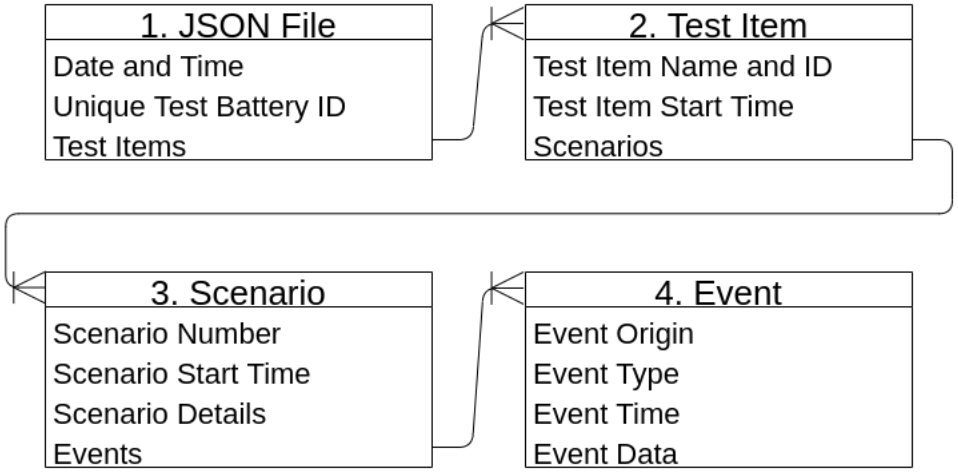


Figure 3.4: Structure of JSON file storing all information of the test being performed. Each of the connections between the tables is a one to many relationship indicating that one table can have many of another table (for example, the Test Item entry can have many scenarios, which in turn can have many events).

viewed in section A.3.

Scenarios are how each of the test items is made variable. The variability is important because of the need for translations and cultural adaptations; it is sometimes only the content shown that needs to be changed (Nampijja *et al.*, 2010). The application contains a database with data detailing scenarios for each test item. By changing the entries in the database, the test items scenarios can be varied, such as changing images to be displayed, words to be read aloud, and the time allowed for a certain task. Each test item has a scenario structure, as each item can be varied in different ways. Therefore the structure of a scenario is set up generically:

instructions - A string denoting the instructions to be read aloud to the participant.

group - The resource item group to be used by the test item denoted by an alphanumerical character.

index - It is the index number used to indicate which resource item should be used as the primary resource item.

rows and columns - The number of rows and columns needed if the test item has a grid.

name - The name of the test item this scenario is applied to.

time - It is an integer number that is used by the test item to control time-related events.

background - The resource item group to be used as a background by the test item denoted by an alphanumerical character.

counter 1 and counter 2 - Integers to help specify custom behaviour in test items.

Resource items are the basic objects used by each test item. If a word/image/description needs to be displayed, it is stored as a resource item in the resource item database. Test item backgrounds and buttons can also be altered by changing the resource items used. Each resource item belongs to a resource group such that test items that require randomness can ask the database for resources of a certain type. The resource groups are defined in section A.4. The structure of a resource item is:

identifier - A unique name that gets logged whenever the resource item is interacted with

word - When a test item requires a word to be displayed or read aloud

image - Link to a displayable image

description - A description to present and read aloud when needed, or to store metadata

group ID - An ID that sorts resource items into groups and it is denoted by an alphanumerical character starting at 0 - 9 and continuing from **a** - **z**

These variables are only used if the test item requires them. In the case where a variable is not applicable, the test item ignores the data in the Scenario data type, and it is stored as a 0, a blank string "", or a Null type.

3.2.2 Origin of Test Items

There are in total eighteen test items, ten language assessment items and eight fine-motor assessment items. Each test item was gathered from literature or a pre-existing standardised test battery where its validity has been tested. Test items were selected based on implementability, and whether or not it would keep its construct validity and use when implemented on a tablet. Furthermore, industry professionals were asked to weigh in on the proposed test items, and alterations were made according to their feedback.

The origin of each test item is essential for the process of refinement and the validity of those test items. As previously mentioned, each of the test items was gathered either from literature or standardised test batteries. Hereafter, short elucidations of each test item's origin follow.

Memory is a crucial domain to measure in the context of measuring language ability, more specifically phonological short-term memory (Conti-Ramsden

and Durkin, 2012) which is one's ability to maintain speech-related information temporarily. The need for measurement is further made evident by the presence of memory test items such as, but not limited to, BAS3's "Recall of objects", Griffiths' "Repeat sentences of 12+ syllables", and CELF's "Number repetition forwards and backwards" and "Recalling sentences". Three test items related to phonological short-term memory are thus included in the tablet assessment, Number Recall, Sentence Recall, and Object Recall.

Similar to the BPVS-III, ROWPVT-4, and PPVT-IV tests, Choose Associated Object and Choose Associated Words tests the participant's receptive language skills. The tests present a participant with a stimulus - which is either a verbal description of something or an object being displayed - and the participant must select the object or word most closely associated with the stimulus. Very similar to Choose Associated Object/Words is the Choose Picture test item, but instead of a single object or word used to describe it, a description is read aloud similar to the EYT tablet assessment's "Not This" test item.

To further measure receptive language skills, DDST and CELF both have test items where the participant is told to follow given instructions and perform an action. In the DDST test item, the participant is given a series of instructions to follow, and the score is calculated from the number of instructions followed, such as "Pick up your shoes and put them on the chair". CELF takes a similar approach where instructions are given to a participant, and how well the instructions are executed is used as the score. Thus the Follow Instructions test item is present in the tablet assessment.

In many of the speech and language assessments performed by speech therapists, having the participant pronounce simple words and listening to the pronunciation of those words is not explicitly a test item. However, the therapist or medical professional administering the test will note problems in the pronunciation. Similar test items can be found in the CELF where a focus on phoneme pronunciation and word structure is measured. Therefore, the Word Pronounce test item is included in the developmental assessment.

As important as it is to measure receptive language skills, it is equally important to measure expressive language skills. Expressive language skill assessments can be found in most assessment batteries, such as DDST and BAS3 where the participant must define words, CELF has an array of test items to assess expressive vocabulary, EOWPVT where the participant must name the object, and Griffiths where picture descriptions are used. Expressive language skill is also measured in the EYT tablet assessment where the participant must name animals he/she/they sees on the tablet. Therefore, a test item Describe Picture is added to assess the participant's language skill.

Present in both DDST and Griffiths, a Give Opposite test item is added to test both receptive and expressive language skill. In practice, speech therapists use opposites, alongside a whole array of other language assessments, to gauge a child's language skill.

In order to measure manual processing speed and manual coordination, finger tapping tasks were added. The tablet test described by Pitchford and Outhwaite (2016) contains such tasks. The participant has to tap as quickly as possible a predetermined amount of times whereby various measures are recorded. Finger tapping can be found in the ZNA as well as the MAND. The finger tapping tasks added to the tablet assessment being designed is Timed Dot Tapping and Rhythmic Dot Tapping. The Timed Dot Tapping measures both manual processing speed and manual coordination, by having the participant tap the dot as fast and accurately as possible. In contrast, the Rhythmic Dot Tapping test item uses the same dot tapping set-up to measure the participant's ability to keep rhythm though motor movements, as rhythmic ability is found to be a predictor of fine-motor skill (Avanzino *et al.*, 2016).

Redrawing objects shown is one of the many ways to measure visuomotor integration and spatial intelligence. Two distinct types are found in known developmental assessments, drawing and copying. Both test items show an object to the participant and require the participant to draw the object, but the difference is that in drawing tasks the object is not present through the entire task but shown briefly. Thus the participant has to draw from memory. Found in DDST, PDMS, and Griffiths, this task was added to the tablet assessment as Draw Object Given.

In classical developmental assessment tests, fine-motor skills can also be measured by the manipulation of objects in the participant's hands. Pick and place test items are used to measure fine-motor as the participant must grasp the object, rotate, and move it over to where the object must be placed. As the manipulation in-hand portion of such tasks would be challenging to measure using a tablet, only the motor planning and rotation part is measured. Many developmental assessments have tasks measuring this pick-and-place and visuospatial aspect of fine-motor, such as Purdue Pegboard Test, BOT-2's transferring pennies and pegboard placing, DDST's raisins in a bottle, PDMS's inserting shapes, ZNA's pegboard placing, MABC's posting coins, and MAND's beads in a box. All these test items require the participant to pick up an object, rotate and move it, and place it in a predetermined hole of sorts. Thus, moving an object, and having to fit it into a hole is added as the Place Object Exactly test item.

As with the Place Object Exactly test item, manipulation of objects is difficult to measure with a tablet assessment. Building and stacking objects on top of each other have been used in various developmental assessments as well. The main focus is on gauging the participant's ability to manipulate objects, spatial intelligence, and motor planning. PDMS has a variety of items requiring the participant to build structures. Likewise, DDST has a test item requiring the participant to build a tower with blocks. Another way to measure motor planning and spatial intelligence is puzzle building, although the puzzle difficulty plays a vital role in the validity. Therefore, the test item Building Object takes the form of a puzzle-like structure to be built by dragging and

placing the puzzle pieces.

In order to measure precision motor skill, and keep the participant engaged, Colour Between Lines test items was added. Present in both PDMS and BOT-2 fine-motor assessment batteries, a metric of how well the participant can stay within lines gives assessors an indication of precision motor skills.

Similar to the construct Colour Between Lines measures, connect-the-dots test items measure precision motor skill as well as motor planning. Again, present in both PDMS and BOT-2, the Connect The Dots test item was added and require the participant to complete a connect-the-dots task.

Both Colour Between Lines and Connect The Dots test items give feedback to the participant in the form of colour or a line being drawn. This feedback helps the participant adapt their behaviour. MABC and BOT-2 both contain tracing path or line test items whereby the participant is required to trace (either with a pen or their finger) a line or trace within a line. Again, precision motor skill is the primary construct being measured, but this time without measuring visuomotor integration (which is tested when visual feedback is given to the participant). Tracing Line/Path is an implementation of such test items but on a tablet. As the Connect The Dots test item already assesses tracing with visual feedback, the Tracing Line/Path test item will not give visual feedback to the participant.

3.2.3 Test Items

3.2.3.1 Number Recall

The goal of this item, seen in figure A.1, is to test the participant's short term memory with regards to numbers. The procedure of the item is to present the participant with a stimulus, wait a predefined amount of time, and then record the response. The stimulus will be a number sequence of variable length being presented to the participant visually (the number is shown on screen) and audibly (the number being read aloud by the tablet). The objective of the participant is to verbally repeat the string of numbers given back at the tablet while it is recording.

This test item uses resource group 3, number objects, and it acquires the entire number object set (0 - 9) from the database upon starting, and the number is displayed using a WordTextView component. The test item uses **counter 1** to determine the number of numbers to insert in a sequence randomly, and the **time** field to determine how long each number is displayed before the next. Both of these variables are used by a timer to control how long the numbers are displayed. The timer starts at the beginning of the scenario, and after **time** seconds it changes the number. This process happens **counter 1** amount of times, and the timer ends. The number to be read and displayed is stored in the resource item's **word**.

3.2.3.2 Sentence Recall

The goal of this item, seen in figure A.2, is to test the participant's short term memory with regards to words. The procedure of the item is to present the participant with a stimulus, wait a predefined amount of time, and then record the response. The stimulus will be a sentence of varied length being presented to the participant visually (the word being shown on screen) and audibly (the word being read aloud by the tablet). The objective of the participant is to verbally repeat the sentence given back at the tablet while it is recording.

The sentence objects group, group 5, is used for this test item. The sentence itself determines the length of the sentence, and it acquires the selected sentence object from the database using the **index** variable and displays it using a WordTextView component. The **time** field is used here as the amount of time before the sentence is no longer displayed after the sentence has been read aloud. A timer is set with the **time** amount, and when the timer ends, the sentence is hidden. The sentence is stored in the resource item's **word**.

3.2.3.3 Object Recall

The goal of this item, seen in figure A.3, is to test the participant's short term memory regarding objects. The procedure of the item is to present the participant with a stimulus, wait a predefined amount of time, and then present a grid of options from which the participant can choose. The stimulus will be a picture of an object. The grid of options will contain buttons with objects on them, and the objective of the participant is to select the correct object shown previously.

Object Recall is varied using the normal objects group, group 1. The object that is shown to be remembered is acquired from the database using the **index** field and is displayed using an ObjectImageView component. The option grid displayed after the stimulus object is done using an AutoTable component and its size is determined by the **rows** and **columns** variables in the scenario details. The database is queried for random resource items from group 1 in order to populate the rest of the options grid. The amount of time between displaying the object and presenting the participant with the grid of options is determined by the **time** field and implemented using a timer that hides the object when the timer ends.

3.2.3.4 Choose Associated Word

The goal of this item, seen in figure A.4, is to test the participant's receptive language. The item's procedure is to present the participant with a stimulus and have the participant choose an option that correlates with the stimulus. In this case, the stimulus will be a picture of an object, and the options will be words on buttons. Each of the buttons, when pressed audibly, pronounces

the word and becomes highlighted. Once the participant is satisfied with the option selected, the DONE button must be pressed to continue.

The scenario dictating the behaviour of this test item uses resource group 1, normal objects. The **index** specifies the stimulus object which is acquired from the resource item's **image** field and displayed using an ObjectImageView component. An AutoTable component is used to display the option and the **rows** and **columns** variables determine the amount of options to choose from. The grid is again populated randomly from the resource group, and only the resource item's **word** is displayed. When an option is selected and highlighted, the **word** is used by the Android Text-To-Speech library to say the word.

3.2.3.5 Choose Associated Object

The goal of this item, seen in figure A.5, is to test the participant's receptive language. The item's procedure is to present the participant with a stimulus and have the participant choose an option that correlates with the stimulus. In this case, the stimulus will be a word, and the options will be pictures of images on buttons. Each of the buttons, when pressed, audibly say what its object is and becomes highlighted. Once the participant is satisfied with the option selected, the DONE button will be pressed to continue.

Similar to Choose Associated Words, this test item uses the normal objects, resource group 1, implements an AutoTable with a size determined by the **rows** and **columns** variables and the **index** field to acquire the stimulus object shown using a WordTextView component. The difference is that instead of using the **image** field as the stimulus and the **word** field as options, it uses the **word** field as stimulus and the **image** fields as options.

3.2.3.6 Follow Instructions

The goal of this item, seen in figure A.6, is to test the participant's receptive language. The procedure of the item is to present the participant with an environment containing actors, instructions will then be read aloud, and the participant's response will be recorded. The environment will be a background image portraying a scene, such as a living room or a kitchen. Actors will be in the form of images overlaid on top of the environment, such as cats, dogs, or people. Instructions will be in the form of "Tap the [colour] [object] in the picture.", or "Tap the [first object] [preposition] the [second object]." Examples are: "Tap the *red* **cat**", or "Tap the **cat** *on top of* the **couch**."

Both background resources, group a, and normal object resources, group 1, are used in this test item's scenarios. The scenario structures a set amount of objects on a background resource and instructs the participant to select a particular object. The background object is acquired using **counter 1** as an index and displayed using an ObjectImageView component. The object to select is acquired using the **index** field and also displayed using an Ob-

jectImageView component. The background resource's description houses the necessary information to place objects on the background. A set amount of coordinate locations are defined, and the test item uses this to populate the image. These coordinates are specified in ratios as to be scale-invariant.

3.2.3.7 Word Pronounce

The goal of this item, seen in figure A.7, is to test the participant's pronunciation. The procedure of the item is to present the participant with a stimulus and then record the response. The stimulus will be a word being presented to the participant visually (the word being shown on screen) and audibly (the word being read aloud by the tablet). The objective of the participant is to repeat the word while the tablet is recording audibly.

The resources used in this test item's scenarios are normal object resources, group 1. The word to be read and repeated by the participant is the resource item's **word** field, and the **index** field determines which one of the resources is used as the stimulus. The word is then displayed in a WordTextView component, and a RecordButton component is stationed next to it to allow for recordings to take place.

3.2.3.8 Describe Picture

The goal of this item, seen in figure A.8, is to test the participant's expressive language. The procedure of this item is to present the participant with a stimulus and then record the response. The stimulus will be a picture of some person or actor performing an action, such as someone tying their shoelaces or waving. The objective of the participant is to describe the picture while the tablet records the response verbally.

The scenario of this test item only specifies the resource group 2, which is the describe objects group, and the **index** of the item to be displayed. The image is acquired from the resource item's **image** field. The image is displayed using an ObjectImageView component accompanied by a RecordButton to be able to record the participant's response.

3.2.3.9 Give Opposite

The goal of this item, seen in figure A.9, is to test the participant's vocabulary and pronunciation. The procedure of this item is to present the participant with a stimulus and then record the response. The stimulus will be a word and picture of an object being presented to the participant visually and audibly (the word being read aloud). The objective of the participant is to give the opposite of the stimulus verbally. Examples include hot/cold, tall/short, big/small.

The scenario specifies the resource group, group 4, and the **index** variable indicates the specific object. The opposite object contains both the word to display and the opposite, which are both logged. The word and image

displayed come from the resource item's **word** and **image** fields respectively. The stimulus word is then displayed in a WordTextView component, and a RecordButton component is stationed next to it to allow for recordings to take place.

3.2.3.10 Choose Picture

The goal of this item, seen in figure A.10, is to test the participant's receptive language. The item's procedure is to present the participant with a stimulus and while presenting a grid of options. The stimulus will be a description of one of the objects in the grid. The grid of options will be buttons with objects on them; the participant will be instructed to select the option that corresponds to the description given.

The scenario of this test item can vary the size of the grid of images displayed by using an AutoTable component by changing the **rows** and **columns** variables and the chosen stimulus from the normal objects group, group 1, using the **index** variable. The stimulus of the chosen object's description in the resource item **description** field will be presented using a DescriptionTextView component and read aloud to the participant.

3.2.3.11 Timed Dot Tapping

The goal of this item, seen in figure A.11, is to test the participant's manual processing speed. The item has a button in the middle of the screen and a timer at the top. The participant will be instructed to tap the button as fast as possible. Once the participant has tapped the button, the timer counts down. Once the time has run out, the tapping button will be disabled.

The scenario can vary the button resource presented, as well as the amount of time allotted for the tapping task. The **time** field defines the count down timer's time, and the **index** selects a button object to display from the dot resource group, group 7. The resource item's **image** field defines the image to be displayed as the button. The button is implemented using a TouchButton component.

3.2.3.12 Rhythmic Dot Tapping

The goal of this item, seen in figure A.12, is to test the participant's visuomotor coordination. The item presents the participant with a button in the middle of the screen and a stimulus. The stimulus will be both visual - border of the screen becomes black and fades to white - and auditory - a beep sound of 1000 Hz played for 300 ms. Between sub-tests, the time between stimulus firing will be varied. The participant must tap the button to mimic the beat of the stimulus. After a predefined amount of taps, the stimulus will stop firing, and the participant must continue with the rhythm until the button becomes disabled, and it is indicated that the test item is complete.

As with the Timed Dot Tapping, the scenario of this test item can vary the button resource. The **index** field determines which object of the dot resource group, group 7, is chosen to be displayed. The **time** field defines the inter-stimulus firing time which will be regulated using a timer and stopped once the desired amount of stimulus aided taps have been made. The **counter 1** field defines the total amount of taps needed to complete the task, and **counter 2** defines the number of taps needed for the stimulus to stop firing - in the case where the participant is to continue with the rhythm for a short while. The button is again implemented using a TouchButton component.

3.2.3.13 Draw Objects Given

The goal of this item, seen in figure A.13, is to test the participant's precision motor control. The item presents the participant with an image to be drawn and space (directly adjacent) to draw the picture. Images will be simple (basic geometric shapes and combinations of these shapes). The objective of the participant is to redraw the image given with their finger accurately.

The test item's scenario only specifies the resource that is used for the stimulus and the resource group where the resource is acquired. The **index** field defines which of the normal object resources in resource group 1, should be used as the stimulus. The **image** field of the resource is then used to acquire the resource's image and displayed using an ObjectImageView component. The PaintView component is set adjacent to the stimulus image and is the location where the participant is instructed to copy the image over.

3.2.3.14 Place Object Exactly

The goal of this item, seen in figure A.14, is to test the participant's visuospatial motor function. The item presents an object and a slightly bigger outline of the object (hereafter referenced as the hole). The participant will be instructed to drag and rotate the object to fit into the hole. Once the participant is satisfied with the placement and rotation of the object in relation to the hole, the DONE button can be pressed to continue to the next test item.

Only the resource to be displayed to the participant and the resource group can be varied using this test item's scenario. The **index** field specifies which resource from the normal objects resource group, group 1, is to be displayed. The resource's **image** field is then used to acquire the image. The hole is displayed using an ObjectImageView component while the moveable object is implemented using a MoveImageView component.

3.2.3.15 Build Object

The goal of this item, seen in figure A.15, is to test the participant's fine-motor precision and motor planning ability. The item presents the participant with a faded depiction of an object and a set of puzzle pieces. The participant

will be instructed to drag and drop the puzzle pieces into the correct location. The objective of the participant is to place the puzzle pieces as accurately as possible onto the faded image.

The test item's scenario specifies the resource to be displayed to the participant and the number of puzzle pieces. The **index** field specifies which resource from the puzzle objects resource group, group 8, is to be displayed. The resource's **image** field is then used to acquire the image. The number of puzzle pieces is specified in the **rows** and **columns** fields of the scenario. The stimulus resource item is implemented with an ObjectImageView component and each piece of the puzzle with a MoveImageView component.

3.2.3.16 Colour Between Lines

The goal of this item, seen in figure A.16, is to test the participant's precision motor control. The item presents the participant with an outlined object, and the participant will be instructed to paint using their finger in order to colour-in the outlined object. The objective is to colour-in the object without painting over the sides or outside of the object.

The scenario of this test item only specifies the resource to be displayed to the participant. The **index** field specifies which resource from the normal objects resource group, group 1, is to be displayed. The resource's **image** field is then used to acquire the image. The components present in this test item are an ObjectImageView to display the stimulus image and a PaintView on top of the ObjectImageView component.

3.2.3.17 Connect The Dots

The goal of this item, seen in figure A.17, is to test the participant's fine-motor precision and motor planning. The item presents the participant with a connect-the-dots object - an outlined or partially completed image with a segment or the entire outline replaces with numbered dots. The participant will be instructed to complete the image by dragging their finger between each dot and connecting the dots in incremental order. The objective is to draw straight lines between each dot in the correct order.

The scenario of this test item only specifies the resource to be displayed to the participant. The **index** field specifies which resource from the connect-the-dots objects resource group, group 8, is to be displayed. The resource's **image** field is then used to acquire the image. Similar to the Colour Between Lines test item, the stimulus object is displayed using an ObjectImageView component with a PaintView component on top of it to allow the participant to draw.

3.2.3.18 Tracing Line/Path

The goal of this item, seen in figure A.18, is to test the participant's fine-motor precision. The item presents the participant with one of three types of lines - straight, curved, or square. The participant will be instructed to trace their finger on the line. The objective of the participant is to trace the path given with their finger as best they can.

The scenario of this test item only specifies the resource to be displayed to the participant. The **index** field specifies which resource from the line and path objects resource group, group 9, is to be displayed. The resource's **image** field is then used to acquire the image. Again, the line/path resource is displayed using an ObjectImageView component with a PaintView component on top, but the PaintView component does not show the lines drawn to the participant.

3.2.4 Data Gathered

The manner in which the data is logged throughout the entire application is by saving all the information in a JSON file for later processing. The file has a generic structure illustrated in figure 3.4. The time that is logged hereafter refers to date and time logged up to a 1000^{th} of a millisecond.

Recordings are stored for the Number Recall, Sentence Recall, Word Pronounce, Describe Picture, and Give Opposite test items in the tablet's file system as audio files as the recording cannot be stored in the JSON file. Therefore, the location of the recording file is stored within the scenario's events log as an event generated by the RecordButton. Accompanying data is also stored, such as the presented stimulus, the scenario's description, and any interaction with the tablet.

The remaining language items - Object Recall, Choose Associated Word, Choose Associated Object, Follow Instructions, Choose Picture - consider only button presses as data. Each of these items requires the participant to select an object/button on the screen. Each selection made is recorded along with the action's timestamp, the object's location, and what it is.

From tapping tasks - Timed Dot Tapping and Rhythmic Dot Tapping - each tap action is recorded. A tap action consists of a time at which the tap occurred and an (X,Y) coordinate on the screen of where the tap took place. Each tap action is counted internally as well. The rest of the fine-motor tasks - Draw Objects Given, Place Object Exactly, Build Object, Colour Between Lines, Connect the Dots, Tracing Line/Path - all record finger-movement action data. Finger-movement action data is how the tablet records the participant tracing their finger on the screen. Every time the participant places their finger on the screen, the action ("Finger Down") is recorded along with the (X,Y) coordinates of the finger on the screen. As the participant moves their finger, the action ("Finger Move") along with the new location the finger

moved to, is recorded. Finally, when the participant lifts their finger off of the screen, the action ("Finger Up") is recorded along with the location where the finger last touched the screen. Each of these actions is recorded as separate events, therefore not constraining the participant to complete the test items in one motion.

Only the finger-movement action data is saved for the Tracing Line/Path test item. As the resources used (the image of the line/path) are available to the analysis pipeline, only the finger-movement action data is needed.

Both the final picture and finger-movement action data are saved for the Draw Objects Given, Connect The Dots, and Colour Between Lines test items. In the case of Draw Objects Given, the picture saved is what is drawn on the canvas adjacent to the given object. Connect the Dots saves the entire picture where the participant has drawn a line to connect the dots. Colour Between Lines also saves the entire screen image. Similar to the recordings, the images cannot be stored directly in a JSON file, but are instead stored in the tablet's file system, and the location is recorded in the JSON file.

Finally, location and orientation data is saved alongside finger-movement action data for Build Object and Place Object Exactly. Place Object Exactly saves the finger-movement action data that is used to drag the object to the hole and saves the final orientation and (X,Y) coordinated. The hole's orientation and (X,Y) coordinates are saved in the scenario block to be able to determine how well the object was placed in the hole. The Build Object test item labels each finger-movement action with the puzzle piece number that was dragged with that specific movement. The original and final (X,Y) coordinates of each of the pieces are saved.

3.3 Data Processing

3.3.1 General Structure and Set-up

The processing algorithm is a set of processing functions that receives as input the generated JSON file, the images, and the recordings. It then iteratively works through the JSON file to determine what test items the test battery consisted of and analyses each accordingly. The output of the program is given as graphs and numerical values which can then be interpreted. The processing is divided into six groups:

- Option Selection - Test item data that consists of the participant selecting an option.
- Placement Accuracy - Test item data where the placement and movement of objects are recorded.
- Tap Error and Time processing - Test item data where the accuracy of tap actions and time-related to tap actions are recorded.

- Tracing Accuracy - Test item data where the accuracy of tracing or drawing a line is recorded.
- Image analysis - Raw images processed and analysed.
- Audio analysis - Raw audio recordings that are to be processed and analysed.

Within the Placement Accuracy, Tap Error and Time processing, and Tracing Accuracy processing categories, distance error will be calculated. Distance error will be the distance between a location the participant has tapped their finger or placed an object, and the desired location as determined by the test set up. The distance error is calculated using two metrics, Manhattan and Euclidean distance. Manhattan and Euclidean distances are used in various processing groups and are both metrics to define the distance between two points, for example $\mathbf{p}$ and $\mathbf{q}$. Both distances are used as Euclidean can provide an overview of the distance error, but Manhattan can indicate an offset in a specific axis. Manhattan distance is the distance between two points on a coordinate system measured at right angles. The Manhattan distance used in the processing of these test items is kept as an array of differences of each of the axes.

$$\text{Manhattan distance } (p, q) = \sum_{i=1}^{n} |p_i - q_i| \tag{3.1}$$

Euclidean distance is the distance between two points in a straight line.

$$\text{Euclidean distance } (p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3.2}$$

3.3.2 Option Selection Processing

The processing of option selection items entails checking that the selected answer is the correct one, and noting how many selections were made before the final answer was chosen. Choose Associated Words, Choose Associated Objects, Choose Picture, Object Recall, and Follow Instructions are seen as option selection items. A stimulus is presented, and the correct option is to be selected.

The processing of one of the test items mentioned above starts by iterating the test item's JSON data and determining the number of scenarios the test item had. Within each scenario's event log is the resource item that is deemed to be correct. The events log is iterated through, and option selections are counted, compared to the correct resource item, and stored along with the time of the action. The final selected option's identifier is compared to the stimulus' identifier, and if it matches, a point is awarded for the scenario being completed. The scenario start time is then subtracted from the final

selection time and stored. Finally, after each scenario has been processed, the average number of items selected before the final item was selected, the average time to the final selection and the number of correctly answered scenarios is calculated.

The proposed performance of a typically developing child is to select the correct answer as quickly as possible, without hesitation. Atypically developing children may take a longer time to process the information presented (instruction given, object shown, word verbally read, or object's description), might not understand the presented information and select the incorrect option, or may hesitate by selecting a few options before deciding on the final selection.

3.3.3 Placement Accuracy

Placement accuracy, in essence, is how accurately the participant placed the object in the specified location or what is the distance from the object's final location to the desired location. The desired location for the Place Object Exactly test item is the hole, which upon rendering into the application logs its location and the random orientation it was given. These properties are logged in the scenario's events log. As the object is being moved, the location and rotation are logged continuously. Therefore, the final location and rotation of the object is the last logged location. Each movement is also logged as an event in the scenario's event log along with the necessary information. Both the Manhattan and Euclidean distances are calculated between the two locations (the hole's location and the object's final location) and the difference in rotation as well. The distance, or placement error, is used to calculate the average of the error for the test item.

A similar process is followed for the Build Objects test item. The test item acquires a puzzle resource which contains an image and splits the image into the predefined amount of tiles (indicated by the rows and columns variables in the scenario's details). Each tile's location is logged in the scenario's events log before being scattered. As each of the objects is being moved, their locations are also logged in the scenario's events log. Therefore, the final location of each object is the location of the last movement event. Again, Manhattan and Euclidean distances are calculated for each tile, and an average is calculated for the scenario (as there may be multiple tiles per scenario). These values are used to calculate the test item's average.

The estimated performance of a typically developing participant is to move, rotate, and place the object in the correct place, thus resulting in a lower distance error. The higher the distance error, the further away the object was placed. This estimation is similar for the Build Object test item. The Build Object test item can also increase the number of tiles, which will increase the difficulty as each tile will be smaller and have less of the overall image to indicate where it should go.

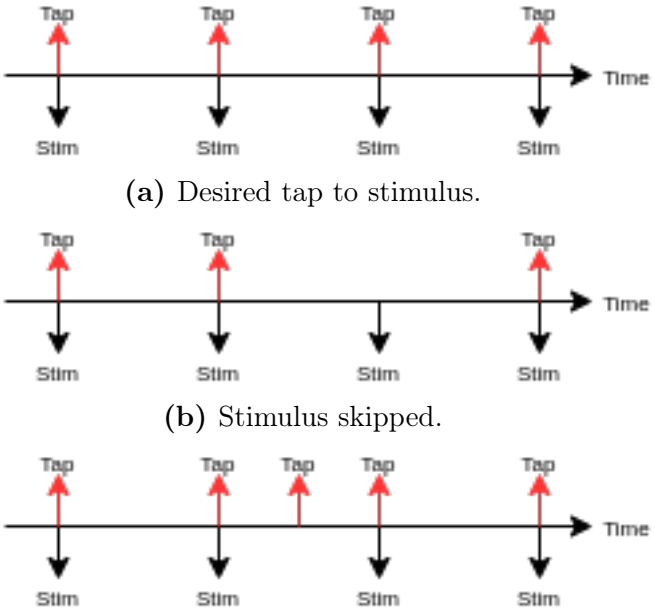
3.3.4 Tap Error and Time Processing

Concerning the tapping tasks, both time and accuracy-related measures are calculated. Along with each tap on the button, the coordinates of the tap and the time are logged. With the rhythmic task, the time at which each stimulus fires is also logged. The two tapping test items are Timed Dot Tapping and Rhythmic Dot Tapping. As a measure of the visuospatial part of fine-motor (hand-to-eye coordination) the accuracy with which the participant can tap the middle of the button is essential. This accuracy can be calculated by measuring tap error.

The tap error is calculated the same for the two test items related to this type of processing, but the timing component is calculated differently. Tap error is the distance, Manhattan and Euclidean, between the tap event and the object's centre. All tap events are logged and stored in each scenario's event log. Furthermore, the object to be tapped logs its location on the screen, and its size upon rendering, and each tap event has the coordinates (X, Y) of the precise tap location. These metrics allow for the calculation of tap errors and the comparison thereof between scenarios. Each tap event's tap error is calculated and used to plot tap error box graphs for both the Manhattan and Euclidean distances. Both metrics of tap error data are then used to calculate an average tap error per scenario and in turn, are used to calculate a test item average tap error. Again, as with the placement accuracy category, less distance error is better. The estimated performance of a typically developing child is to tap the button more centrally than an atypically developing child.

Concerning timing in the Timed Dot Tapping test item, the time between taps is of importance. The time each tap event happens is logged in the scenario's event log. The algorithm iterates through each of the tap events recorded and subtracts the time of each tap event from the prior event's time results in the time between taps. The overall process results in $n-1$ inter tap times, where n is the number of taps that took place during the scenario. The inter tap time is plotted in order to see the inter tap time variance per scenario and then used to calculate a scenario average. This average inter tap time for each scenario is used to calculate an average test item inter tap time. Having short and consistent inter tap time (thus tapping quickly and consistently) is an indicator of good motor timing ability and therefore good overall fine-motor ability. The amount of taps in the allotted time is the final measure of the Timed Dot Tapping test item, whereas the Rhythmic Dot Tapping test item has a set number of taps to be performed, determined by **counter 1** in the scenario details.

The time measurement component of the Rhythmic Dot Tapping test item is to determine how well the participant mimicked and kept the rhythm. It is determined by calculating the time between every tap and its nearest stimulus and determining the number of errors present, which is illustrated in figure 3.5. The number of errors is found by comparing the number of stimuli to be



(a) Desired tap to stimulus.

(b) Stimulus skipped.

(c) Too many taps inbetween stimuli.

Figure 3.5: Illustration of the desired outcome, a stimulus being missed, and a miss tap between stimuli.

tapped (determined by **counter 2**) and the number of stimuli that have fired since the first tap. The tap event's time is compared to the previous and next stimulus' time (the stimulus before and after the tap event in the event's log of the scenario), and for each tap, a tap timing error is calculated. Furthermore, inter tap time is calculated the same as with Timed Dot Tapping, and plotted to acquire the inter tap time variance graph. The tap timing error and inter tap time are both averaged for the scenario. These averages are then averaged again over all scenarios for a test item average. The number of errors made per scenario is averaged as well, giving a test item average.

3.3.5 Tracing Accuracy

Tracing accuracy analysis applies to the Connect The Dots and Tracing Line/-Path test items. The standard or classical way to measure tracing accuracy (as done by occupational therapists in a face to face evaluations (Rhode, 2019)) is to let the participant draw lines with a pen (on a connect-the-dots image or to trace a line) and use a stencil to measure the deviation of the line from the desired line. This process is similar to how the tracing accuracy is processed and calculated.

Each connect-the-dots resource item has in its description field the location of all the points. These locations are in the form of a number between 0 and 1, which makes the location scale-invariant, as the dot's location coordinates would scale with the size. Points are indicated in figure 3.6 as green circles

and labelled A, B, and C hereafter referred to as points. The preprocessing process consists of getting the equation for the line segment between all point pairs. A point pair is two consecutive points between which the participant has to draw a line. The data acquired is a set of coordinates depicting the finger paths drawn by the participant. As the participant places their finger on the screen, the application registers the coordinates where the finger is placed. Each finger movement on the tablet (such as drawing a line) prompts the application with a new set of coordinates to register. The application then checks to see if the new coordinates are far enough away from the previous coordinate set to indicate a finger movement. If it is deemed far enough, the new coordinates are registered as a finger path data point, but if not, then the new coordinates are ignored (as holding one's finger on the tablet will continuously generate touch events). An entire line drawn by the participant is thus a set of coordinates with some necessary information such as what type of action prompted the coordinate pair to be registered (a finger movement or a finger placement). The line drawn by the participant is seen as a blue line and consists of numerous coordinate points registered by the application. In order to simplify the further explanation, the three red dots indicated in figure 3.6 and labelled 1,2, and 3 will be taken as samples and used as examples, hereafter referred to as dots. Each of the dots then has to be assigned to a line segment. The next step is to determine which line segment to use to calculate a particular dot's error distance. Assigning a segment is done by using the three nearest points to any given dot and assigning that dot to the line segment between the closest two consecutive points. In the case of dot 1, the joint euclidean distance from dot 1 to points A and B is less than that of the distance from dot 1 to points B and C. Once each dot from the participant's finger path has been associated with a line segment, the deviation distance needs to be calculated. The error is the perpendicular distance from each dot to its assigned line segment. If no perpendicular line passes through the dot (in the case of dot 3), the error is calculated as the distance to the nearest point belonging to its assigned segment. For example, the distance between point C and dot 3 will be used.

With the error for each dot calculated the average error per line segment and per scenario can be calculated. This in-turn will be used to calculate the average error per test item.

Similarly, the error for the Tracing Line/Path test item is also calculated as the deviation from the desired line. In this case, the desired line is the line on an image, which simplifies the process to finding the nearest pixel belonging to the line from each of the finger path dots. The pixel search operates iteratively around the dot's coordinate, checking in an increasing radius every pixel around the coordinate until a line/path pixel (indicated by a black pixel, as the line is black) is found. The Euclidean distance is then calculated between the dot and the nearest line pixel. Also similar to the Connect The Dots test item, the average error per scenario is calculated,

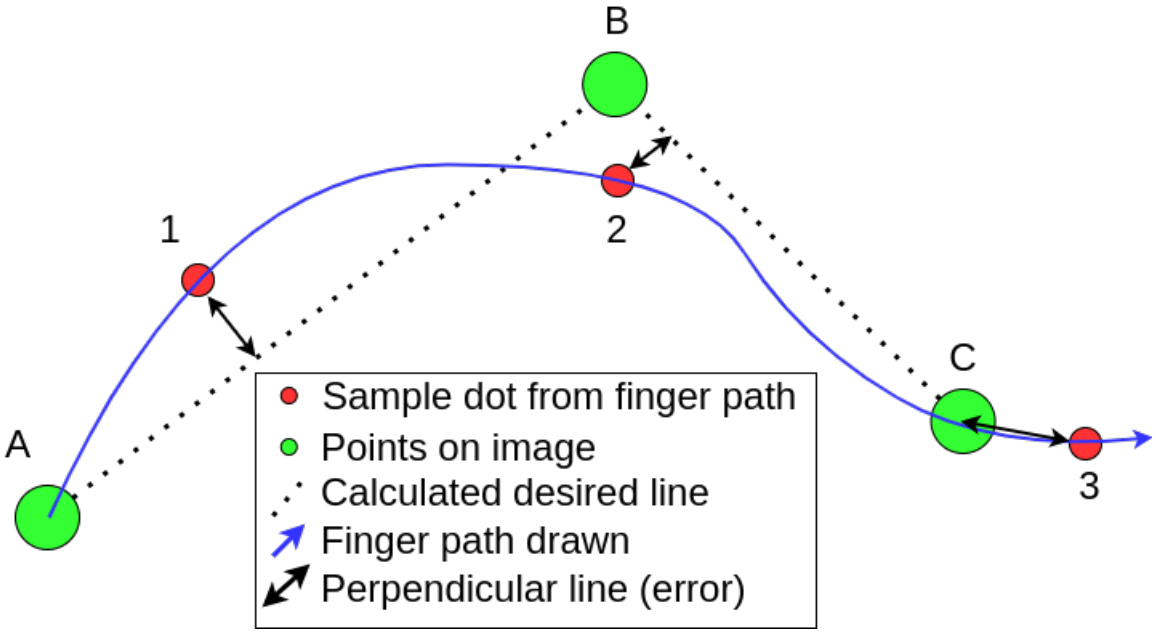


Figure 3.6: Tracing accuracy illustration with legend.

which in turn is used to calculate a test item average.

3.3.6 Image Analysis

This category of analysis pertains to the Draw Objects Given and Colour Between Lines test items. The data acquired from the latter test item is an image taken before and after the participant has coloured-in the given image, hereafter referred to as pre- and post-image, respectively. The criteria for this test item is to colour-in every pixel within the given outlined object without colouring over the object's borders or outside the outline. The performance here is measured in percentage of error pixels present. An error pixel is a pixel inside the object that is still uncoloured and a pixel that is coloured-in but should not be - such as the black pixels of the border and the white pixels outside the object's outline.

This test item has both a pre- and post-image saved to determine which pixels are error pixels and counting them. By comparing modified versions of the images, this problem can be reduced to counting the number of pixels present on the screen. Boolean operations, in this context, are the comparison of bits using specific operations. If certain information is extracted out of the images in the form of a boolean map (a 2D array the size of the image, but having only 1s and 0s), then boolean operations can be used to compare these images. The pieces of information that can be extracted to a boolean map are the borders - acquired from the pre-image, and the user painted pixels - from the post-image.

First, both pre- and post-images are greyscaled, which is the process of converting an image from a colour scale (typically RGB, or Red-Green-Blue which is three 2D arrays, one for each colour) to grayscale (a single 2D array with values ranging from 0 indicating black to 255 indicating white). Greyscaling is done by assigning ratios to each of the colour channels, in this case, a third each. Each colour value is multiplied with the ratio and added up to equal the grayscale value. From here, an image segmentation process is applied to sort the pixels into one of three bins: black (usually pixels that form part of the border), white (uncoloured pixels) and grey (pixels that have been coloured-in). The segmentation process loops through the image, checking each pixel value and assigning them to the appropriate category. Values between 0 - 63 are assigned to category 1, between 64 - 191 are assigned to category 2, and between 192 - 255, are assigned to category 3. These values were chosen by dividing the range of values into four blocks. Therefore, all values close to 0 are sorted into category 1 (border pixels), all values close to 255 are sorted into category 3 (white pixels), and anything in-between is sorted into category 2 (coloured-in pixels). Category 2 receives an expanded range as the grey values of the coloured-in section may vary (depending on what colour the participant chose to use), and category 1 and 3 receive a buffer to ensure all border/white pixels are captured.

Comparison operations can be performed on these modified images if they are converted to boolean maps. In order to acquire the user-painted boolean map, the pre- and post-images are compared with a "DOES NOT EQUAL" or "!=" operator. This operator will result in a boolean value of 1 (meaning True) where the two image's pixels are not the same and resulting in a boolean value of 0 (meaning False) where the two image's pixels are the same. As the only difference between the pre- and post-images are the pixels coloured-in by the participant, the resulting boolean map will indicate which pixels were coloured-in, and will hereafter be referred to as boolean map 1. The next step is to separate the inside of the outlined image (the area that needs to be coloured-in) from the border and outside (the area that should not be coloured-in). This process is started by acquiring a boolean map indicating the border. The modified pre-image contains only two of the three categories from the previous segmentation operation, and so a threshold comparison is used to extract the border. A threshold comparator, which in essence checks if a given pixel's value is lower or higher than a threshold, is used to determine which pixels are dark enough (having a value close to 0) to be a part of the border. This thresholding results in a map with 1s indicating the border and 0s indicating everything else. Next, the pixels outside of the object's border need to be identified. As an image is a 2D array, it consists of a set of rows and columns. A row would then be a slice out of the image from left to right, indicated on figure 3.7. By iterating through each row, each pixel value is compared to determine whether or not it is a part of the border (black pixel). The first of the border pixels (indicated in figure 3.7 with red arrow 1) would

indicate the start of the object's border and therefore any pixel before that is considered outside the border. Similarly, the last pixel (indicated in figure 3.7 by red arrow 2) would mark the end of the border, and any pixel after that is considered to be outside the object. The order of processing is flipped, and the same operation is performed by iterating through columns instead of rows. The result is a boolean map with 1s (boolean true) indicating the border and every pixel outside of the border (everything that should not be coloured-in), and 0s indicating the inside of the object (everything that should be coloured-in by the participant) and will be referred to as boolean map 2.

Finally, the two boolean maps are used to determine the number of error pixels. The operation that gives rise to the final map is the XNOR of the two maps. The XNOR operation is a combination of an XOR and a NOT operation. The NOT operation changes 1s to 0s and 0s to 1s in the boolean map given. The XOR operation, or exclusive OR operation, results in a 1 if, and only if, just one of the compared booleans has a value of 1. Therefore, the XNOR operation results in a 1 if both values are 1s, or both are 0s, and results in a 0 if either one of the values being compared has a value of 1. Thus if map 2 has a value of 1, indicating that the specified pixel should not be coloured-in, and map 1 has a value of 1, indicating that it has been coloured-in, the XNOR would result in a 1, indicating that it is an error. If map 2 has a value of 0, indicating that the pixel should be colour-in, and map 1 also has a value of 0, indicating that it has not been coloured-in, then the XNOR operation would result in a value of 1, indicating that that specific pixel is an error pixel. The final boolean map has values of 1 indicating error pixels, and each of these is counted to acquire the number of error pixels. The number of error pixels, coupled with the total number of pixels in the image (width of the image multiplied with the height of the image), are used to calculate the error percentage for each scenario. The error percentage, in turn, will be used to illustrate the error per scenario and calculate the total average error for the test item.

The Draw Image Given test item gives two images, the stock image that is displayed to the participant and a drawn image that the participant drew to copy the stock image. Determining how well the participant drew the stock image is an image similarity problem, how similar is the drawn picture to the stock picture. Six methods are used to evaluate how similar the two images are in order to show where specific methods work and where some fall short. Each similarity score is a percentage of how well the participant copied the stock image. The score is scaled and offset by having the stock image compared to itself with each metric, resulting in the 100% score mark, and compared to a blank image where nothing is drawn, resulting in the 0% score mark.

There are two sub-fields in the field of image verification (which is whether two images are similar or not), image embeddings and metric learning methods. The former learn a robust and discriminative descriptor with which to represent the image as a feature vector. The latter learns a distance metric

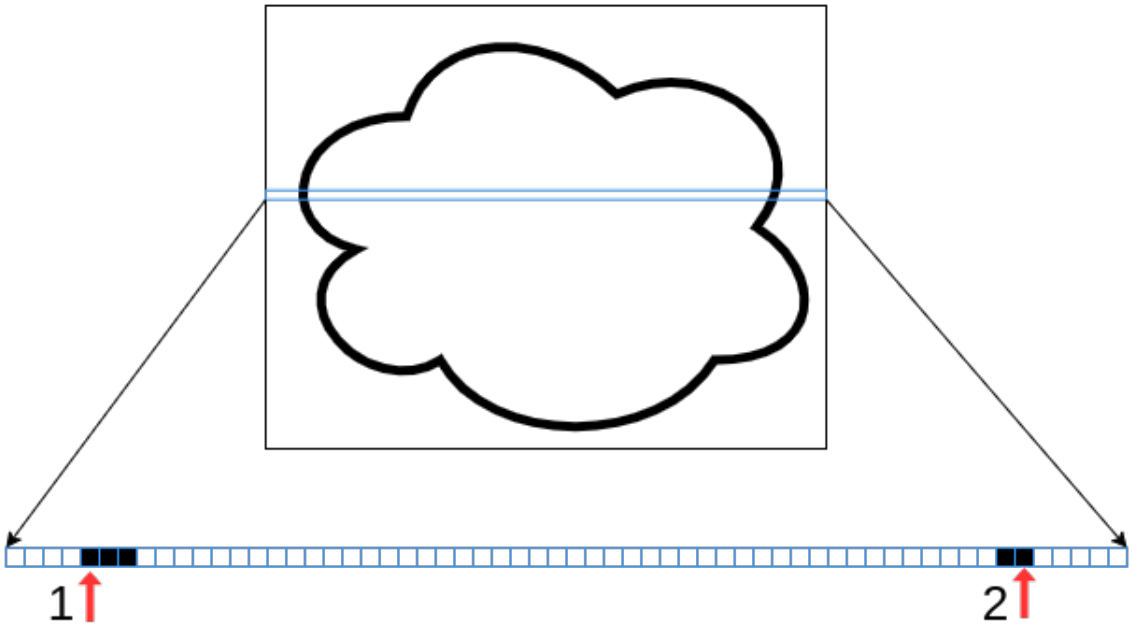


Figure 3.7: Border pixels, indicated with red arrows, are found by iterating through each row in an image, and saving the first and last border (black) pixel found.

from labelled training samples in an embedding space (Appalaraju and Chaoji, 2017). As there is currently a lack of training data to train such a distance metric, image embeddings are used alongside standard distance metrics. Where training is required, in the case of the machine learning approaches based on convolutional neural networks (CNNs), a pre-trained model is used.

The participant, mostly preschool children, are not constrained or guided when drawing the image. Not enforcing constraints results in difficulty with some similarity methods that calculate a pixel per pixel difference distance instead of looking at the image as a whole to determine the similarity. Nevertheless, the sum of squared differences (SSD), cosine similarity (CS), and Hausdorff distance (HD) metrics are included to illustrate the difficulty with scaled-down or translated images. Used in many image-based search algorithms to calculate image similarity, a Scale Invariant Feature Transform (SIFT) (Rey Otero and Delbracio, 2014) algorithm is used as well. The last two metrics used are a CNN based feature extraction and a machine learning-based image similarity application programming interface.

SSD can be described using the following:

$$SSD(I, J) = \sum_{i=1}^{N} (I(i) - J(i))^2 \tag{3.3}$$

where I and J denote the two images being compared flattened to a 1-dimensional array of pixel values and N the number of pixels present. SSD results in a value greater than 0 as a measure of dissimilarity, where 0 is precisely similar

and the larger the value becomes the more dissimilar the two images are. Cosine similarity is calculated between two 1-dimensional arrays, similar to SSD, and is described as:

$$CS = \frac{\mathbf{A} \cdot \mathbf{B}}{||A||_2||B||_2} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \tag{3.4}$$

where **A** and **B** denote the two image arrays being compared, and N denotes the number of pixels present in the images.

Hausdorff distance is also used in calculating the similarity of two images and image matching. The Hausdorff distance is the maximum value of a two way directed Hausdorff distance calculation performed. It equates to each point of A finding its closest neighbour from B, and the most mismatched point of A determines the value of $h(A, B)$ (Ruckligde, 1996).

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{3.5}$$

The scale-invariant feature transform algorithm first published by Lowe (1999) detects and describes features locally in images. The SIFT algorithm is invariant to scaling, rotations, and translation. These characteristics are desirable as the participants can vary the scale/orientation/location of the drawn image (the participant may draw the object a bit smaller, off to one side, or rotated a bit). It mainly consists of detection of key points and extraction of descriptors of those key points. These descriptors are then compared among images to find an image that matches the best. According to Rey Otero and Delbracio (2014), the SIFT algorithm can be divided into eight steps:

- Compute the Gaussian scale-space
- Compute the Difference of Gaussians (DoG)
- Find candidate key points utilising 3D discrete extrema of the DoG
- Refine candidate keypoints location with sub-pixel precision
- Filter unstable keypoints due to noise
- Filter unstable keypoints laying on edges
- Assign reference orientation to each keypoint
- Build keypoint descriptors

CNNs are a type of machine learning algorithm (or neural network derivative, see section B) that receives as input an image and results in a feature vector of the image, for more information, see section B.3. The extraction of

this feature vector is done by having a set of kernels (filters) learn which characteristics are essential during training. The kernels are essentially matrices (of predetermined size) that are cross-correlated with the image's pixel values by sliding it like a window over the image. Numerous configurations of CNNs are used today, each with their advantage, and can have additional layers such as pooling and activation layers. Typically, a fully connected neural network receives the feature vector produced by the CNN as an input and is trained to classify an image according to its feature vector. In this specific case, the feature vector itself is used. The feature vector of the stock image and the drawn image are produced, and then the distance between the two is calculated. This distance is, as previously mentioned, scaled and offset using the distance from the given image to itself, and the given image to a blank image.

A pre-trained model is used, and images are given as inputs. The pre-trained model selected is ResNet-152, further described in section B.5. This model performs better than most models and is not as large as some models (He *et al.*, 2016*b*; Anwar, 2019). The size of the model is important as it has a direct correlation with the amount of time it takes to evaluate an image. ResNet-152 takes as input an image of size $227x227$ with three channels (red, green, and blue). The images acquired from the tablet test may vary depending on the size of the tablet but will be automatically resized to the required size. The output of this network is a feature vector containing a 1000 entries. The Euclidean distance between the two vectors will be used similarly to how the other metrics are compared.

Lastly, DeepAI's Image Similarity API is used as the final metric. The API allows two images to be uploaded and compared, returning a numerical distance metric. Each of the metrics mentioned is used to calculate a score as to how well the drawn image mimics the stock image. Each value is represented as a similarity percentage which is scaled and offset by comparing the stock image to itself (equating to 100%) and to a blank image (equating to 0%). The values are tabled and plotted for each scenario. In turn, each of the metrics' results are used to calculate a metric average for the test item.

All six these measures will be displayed as a percentage similarity score comparing the drawn image to the stock image.

3.3.7 Audio Analysis

In order to do autonomous analysis on audio and language assessment, an automatic speech recognition (ASR) system is needed. This system takes as input an audio file and gives an array of possible transcriptions for the audio file. Before any score or analysis can be done on what the participant has said, the speech recognition system needs to produce transcriptions, and these transcriptions, in turn, will be used to process whether or not the participant completed the task successfully. Two methods of speech recognition are used to acquire transcriptions, DeepSpeech 2 (Amodei *et al.*, 2016)(which was built

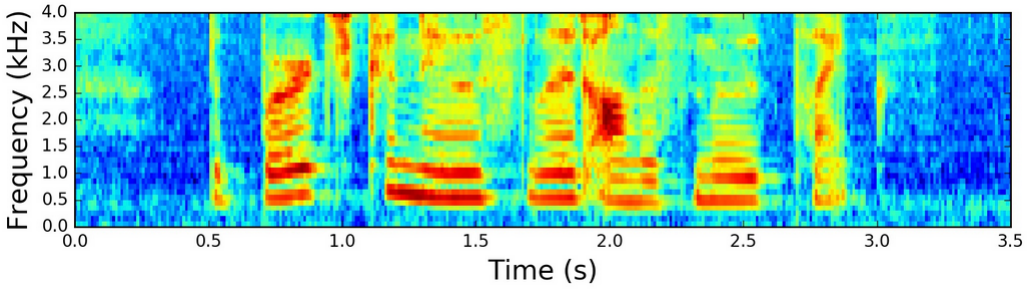


Figure 3.8: A mel spectrogram (Fayek, 2016)

according to the architecture and trained on a dataset LibriSpeech (Panayotov *et al.*, 2015)) and Google's Speech-to-text API, as they are considered the two most popular end-to-end speech recognition models to date.

DeepSpeech 2 uses a convolutional neural network (CNN) and a recurrent neural network (RNN, see section B.4) to probabilistically estimate which letter of the alphabet corresponds to which segment of the audio file. This predication is then correlated to a transcription using Connectionist Temporal Classification (CTC, Graves *et al.* (2006)). CTC is an algorithm that calculates a probability of Y (an output) given X (an input). The input for this specific case is predictions on what letter is said for every time segment of audio. Each prediction is assigned a probability, and the most probable combination of letters is selected as a result. The first step is converting the raw audio file into a mel spectrogram as seen in figure 3.8, which in essence is a depiction of the sound. A mel spectrogram is a visual representation of an audio signal's frequency intensities over time, converted to the mel scale. The mel scale, first proposed by Stevens, Volkman, and Newmann in 1937, is a unit of pitch where equal distance in pitch sounds equally distant to a listener. To acquire a spectrogram, the process of fast-Fourier transforms need to be applied to the audio for each time segment. The results are an array of frequency intensities that are plotted over time, which is then converted to the mel scale.

The essential model of DeepSpeech2 is a combination of two main neural networks, one for analysing the mel spectrograms and extracting unique features from it and one to interpret and predict which combination of those features results in which letter is pronounced. The former is a residual convolutional neural network (ResCNN, He *et al.* (2016*a*)) and the latter is a bidirectional recurrent neural network (BiRNN, Schuster and Paliwal (1997)), both derivatives of a CNN (for more information see section B.3) and a RNN (for more information see section B.4) respectively. Finally, the CTC receives the suggested letters from the BiRNN and forms the full transcription. ResCNNs differ from standard CNNs by having skip connections, or residual connections, nestled into the network. Skip connections allow the output of one layer in the network to skip a layer and form part of the input to the layer after that. Recurrent neural networks are a type of neural network where the nodes

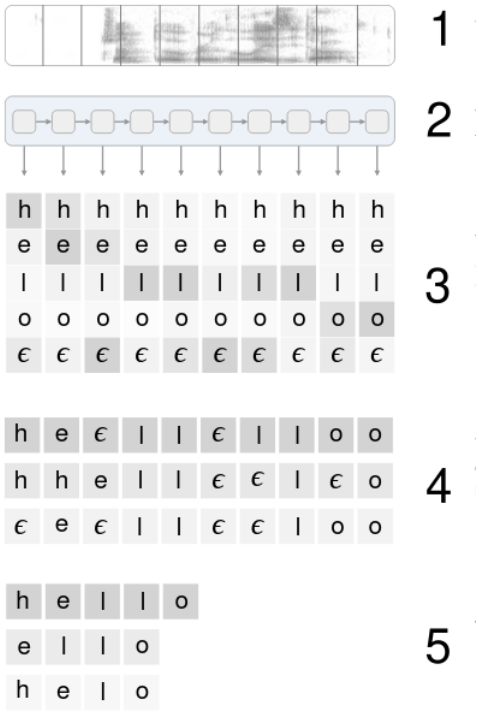


Figure 3.9: Visual representation of CTC in a speech recognition model (Hannun, 2017)

and their connections form a directed graph allowing it to exhibit temporal dynamic behaviour. In essence, it allows the network to look and remember what came before the current piece of data in order to make more accurate predictions. A bidirectional RNN is not only able to look at what piece of data came before the current one but can also look ahead, to make a more accurate prediction. Being able to look both directions is important in speech processing as single slices in time which contain segments of letters being spoken are difficult to predict, but when looking at what is said next and what is said before, the prediction can be made more reliable. CTC receives input sequences, X, that are acquired from the BiRNN where $X = [x_1, x_2, \ldots]$ and outputs a sequence, Y, such as a transcript where $Y = [y_1, y_2, \ldots]$. In figure 3.9 the process is visualised. In step 1, a CNN extracts information from an audio file transformed into a spectrogram. The RNN network receives this extracted information and calculates a probability array of what each time segment could be, seen in step 3. The ϵ character introduced by the use of the CTC algorithm coincides with a blank character and is used when the network thinks a time segment contains no notable information. The CTC then calculates the probability of different sequences in step 4, and this results in a distribution of outputs in step 5. Alongside the DeepSpeech2 model, each audio file is sent to the Google Speech-to-text API, which returns an array of possible transcriptions and probabilities of each.

All five test items require these models to transcribe the audio, but only four out of the five test items have a given audio stimulus to repeat. Give Opposite, Word Pronounce, Sentence Recall, and Number Recall all have a set stimulus with which to compare the participant's transcribed audio. Once the speech recognition has given a transcript, the metric to calculate a score for how well the participant performed will be word error rate (WER) and character error rate (CER) (Seljan and Dunder, 2014). WER will assign a point if a word corresponds correctly to the word needed and none if the word is incorrect. WER can be a harsh metric as simple mispronunciations of specific phonemes can result in a word not being counted. CER, a more lenient metric, checks every character that matches and similar to WER will assign a point for every correct character pronounced but results in a 0 for every character missing or incorrect. With both WER and CER metrics for each scenario, a test item average WER and CER is calculated.

The Describe Picture test item's score will be calculated by counting the number of correct keywords spoken. This measure is similar to how speech therapists assess expressive language, as well as the EYT's expressive language assessment. Each picture to be described will have three keywords that need to be spoken. In the example picture of figure A.8 where a boy is sitting and reading, the keywords would be boy/man, sitting/sit, reading/read. The transcription is searched for these keywords, and a point is awarded for each one. This metric results in each scenario having a score out of three, which is plot to give a per scenario performance graph. The final metric is an average score for the test item.

Chapter 4

Results and Findings

4.1 Introduction

The results presented here are to assess whether or not the data gathering and data processing applications work as intended. Each of the processing categories' results will be given in their sections, explaining the set-up and input used to generate the results. Details of the scenarios, which describes the proceeding of the test item, are stated, the number of scenarios per test item, the number of times each test item was completed, and how interaction with each test items was premised. Two attempts were made for every test item, except test items in the option selection category. The two attempts vary in intended performance and are used to illustrate the higher and lower spectrum of performance for each test, as foreseen by the researchers. All scenarios are consistent across each of the attempts.

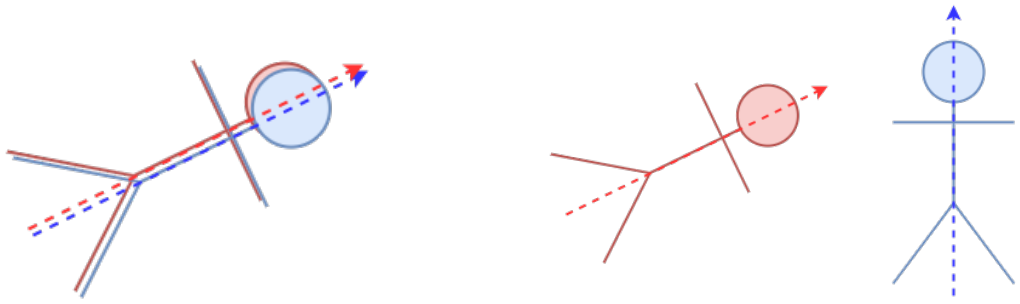
4.2 Option Selection

The test items in the option selection processing category are Choose Associated Words, Choose Associated Objects, Choose Picture, Object Recall, and Follow Instructions. Only one attempt was made for each of the test items in this category as each scenario's performance was varied. Each scenario was completed with a different objective and are as follows: the first scenario was a quick correct answer, the second was a long correct answer, the third was a quick incorrect answer, the fourth a long incorrect answer, and in the final scenario multiple options were selected, but the last option selected was the correct answer. The scenarios also varied the objects being displayed to the participant for all five test items.

The test item's results are shown in table 4.1.

Table 4.1: Option Selection processing of each of the five option selection test items. Three metrics are displayed, whether or not the final answer was correct, the number of selections made during the scenario, and the time it took to make the final selection (in milliseconds). These three metrics are averaged across scenarios to give the participant a test item score.

	Correct	**# of selections**	**Time (ms)**
Choose Associated Objects			
Scenario 1	True	1	1115.0
Scenario 2	True	1	6186.0
Scenario 3	False	1	1502.0
Scenario 4	False	1	5387.0
Scenario 5	True	3	4594.0
Average	60.0%	1.4	3756.8
Choose Associated Words			
Scenario 1	True	1	1566.0
Scenario 2	True	1	5009.0
Scenario 3	False	1	2069.0
Scenario 4	False	1	5937.0
Scenario 5	True	3	5103.0
Average	60.0%	1.4	3936.8
Choose Pictures			
Scenario 1	True	1	1782.0
Scenario 2	True	1	4869.0
Scenario 3	False	1	1608.0
Scenario 4	False	1	4360.0
Scenario 5	True	3	5563.0
Average	60.0%	1.4	3636.4
Object Recall			
Scenario 1	True	1	3603.0
Scenario 2	True	1	8878.0
Scenario 3	False	1	3446.0
Scenario 4	False	1	8243.0
Scenario 5	True	4	11739.0
Average	60.0%	1.6	7181.8
Follow Instructions			
Scenario 1	True	1	2139.0
Scenario 2	True	1	5204.0
Scenario 3	False	1	1320.0
Scenario 4	False	1	6806.0
Scenario 5	True	3	6060.0
Average	60.0%	1.4	4305.8



(a) Representation of correct way to place object on hole (Attempt 1).

(b) Representation of incorrect way to place object on hole (Attempt 2).

Figure 4.1: Correct and incorrect attempts for the Place Object Exactly test item where the red stickfigure represents the hole and the blue stickfigure represents the object to be moved. The red and blue dotted lines indicate the orientation of the object, and the difference in degrees between the two lines' orientations is the rotation error.

4.3 Placement Accuracy

The test items in the placement accuracy processing category are Place Object Exactly and Build Objects. Both test items require the participant to move an object and place it in the correct place and orientation. Five scenarios are recorded for both test items. Attempt 1 for both test items was seen as the "good" attempt and consisted of placing the objects in the correct positions and orientations as best possible to simulate a participant that can perform well on the test item. Attempt 2, which was seen as the "bad" attempt, placed the objects at random locations on the screen and rotated them randomly as well to simulate a participant that is not able to place the objects in their desired locations. The Place Object Exactly test item's scenarios only varied the object being displayed, and a "good" and "bad" attempt can be seen in figures 4.1a and 4.1b respectively. The five scenarios of the Build Object test item varied the dimensions of the object to be built (the number of rows and columns to divide the puzzle image into, thus also varying the number and size of each puzzle piece). Varying the dimensions increases the difficulty as more pieces have to be moved to the desired locations and the puzzle pieces are smaller containing a smaller region of the object being built. Two examples thereof can be seen in figures 4.2b and 4.2c having four big pieces and sixteen smaller pieces. Figure 4.2b illustrates a near perfect attempt, 4.2d illustrates a bad attempt, and figure 4.2c illustrates a good attempt.

The placement errors of the Place Object Exactly test item are found in table 4.2, containing each scenario's mean and standard deviation placement error. Similarly, the placement errors for the Build Object test item are found in table 4.3, also having each scenario's mean and standard deviation.

Table 4.2: Results from Place Object Exactly test item where distances (Manhattan X and Y, and Euclidean) are in pixel distances and rotation is in degrees.

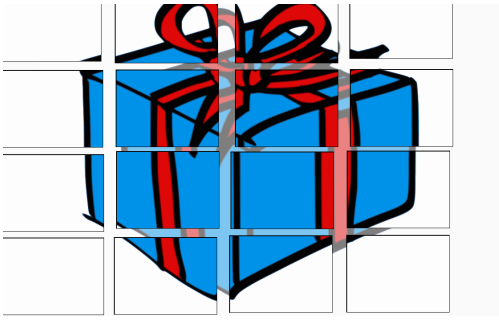
	Manhattan X	**Manhattan Y**	**Euclidean**	**Rotation**
Attempt 1				
Scenario 1	2.070	41.987	42.038	3.803
Scenario 2	2.012	3.003	3.614	5.178
Scenario 3	1.992	32.007	32.069	0.351
Scenario 4	0.039	0.000	0.039	0.172
Scenario 5	2.031	39.014	39.067	6.230
Mean	1.629	23.202	23.365	3.147
STD	0.795	18.037	17.917	2.479
Attempt 2				
Scenario 1	757.969	152.007	773.061	141.000
Scenario 2	667.031	139.014	681.363	42.000
Scenario 3	718.008	96.997	724.530	6.000
Scenario 4	644.023	44.995	645.593	26.000
Scenario 5	753.008	61.011	755.475	141.000
Mean	708.008	98.805	716.004	71.2
STD	45.594	41.891	46.997	58.122

Table 4.3: Results from Build Object test item where distances (Manhattan X and Y, and Euclidean) are in pixel distances and the amount of pieces each scenario had.

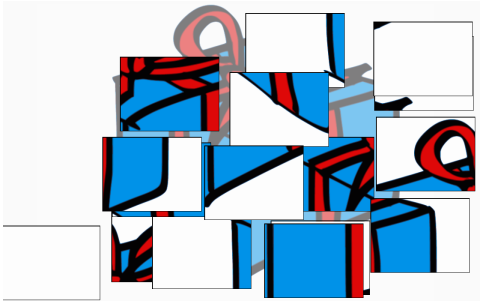
	Manhattan X		**Manhattan Y**		**Euclidean**		**# Pieces**
	Mean	STD	Mean	STD	Mean	STD	
Attempt 1							
Scenario 1	2.020	0.704	4.739	0.423	5.193	0.488	4
Scenario 2	1.676	0.751	4.514	1.991	5.087	1.354	6
Scenario 3	1.975	1.558	5.837	3.436	6.212	3.690	6
Scenario 4	1.114	0.568	2.893	1.093	3.199	0.944	9
Scenario 5	2.333	1.893	2.052	1.714	3.386	2.170	16
Attempt Avg.	1.823	1.095	4.007	1.731	4.616	1.729	-
Attempt 2							
Scenario 1	433.496	213.244	181.507	106.789	486.212	203.317	4
Scenario 2	335.837	256.325	164.507	46.409	392.748	231.200	6
Scenario 3	491.922	368.248	179.495	113.774	582.737	288.400	6
Scenario 4	232.216	145.730	184.998	112.239	320.949	137.750	9
Scenario 5	371.719	219.026	190.140	116.836	435.254	215.650	16
Attempt Avg.	373.038	240.514	180.129	99.209	443.580	215.263	-

(a) Stock image to be reconstructed with puzzle pieces.

(b) Representation of how four puzzle pieces would be placed correctly (attempt 1).



(c) Representation of 16 piece puzzle almost complete, puzzle pieces are in correct region.



(d) Representation of incorrectly placed puzzle pieces (attempt 2).

Figure 4.2: Build Object test item "good" and "bad" examples with varying puzzle dimensions.

4.4 Tap Error and Time Processing

The test items in the tap error and time processing category are Timed Dot Tapping and Rhythmic Dot Tapping. Both test items contain five scenarios each. The Timed Dot Tapping test item's scenarios vary in the amount of time given to the participant to tap the button, starting at 10 seconds with the first scenario and increasing in increments of five seconds with the fifth scenario having 30 seconds.

The main difference between scenarios for the Rhythmic Dot Tapping test item was the time between stimuli or rhythm component. Table 4.6 houses timing-related results for the Rhythmic Dot Tapping test item. The inter-stimulus time for each of the five scenarios was as follows: 500 ms, 1000 ms, 1500 ms, 2000 ms, 2500 ms. Each scenario has 20 taps that are done with a rhythm, and then a remaining ten are done without a rhythm.

The tapping results for the Timed Dot Tapping test item are found in table 4.4, and for the Rhythmic Dot Tapping test item in table 4.5. Examples of "good" and "bad" attempts with regards to tapping accuracy can be seen in

Table 4.4: Average tapping error per metric, per scenario for Timed Dot Tapping attempts in pixels.

	Manhattan X		**Manhattan Y**		**Euclidean**	
	Mean	STD	Mean	STD	Mean	STD
Attempt 1						
Scenario 1	8.655	5.182	8.397	6.689	13.449	6.012
Scenario 2	14.818	10.581	16.955	12.250	24.570	12.860
Scenario 3	21.058	17.854	38.580	19.671	45.503	23.812
Scenario 4	8.417	6.287	14.371	9.785	17.802	9.785
Scenario 5	10.348	8.096	14.884	10.145	19.279	11.198
Attempt Avg.	12.659	9.600	18.637	11.708	24.121	12.733
Attempt 2						
Scenario 1	48.643	32.665	50.071	30.965	71.849	41.675
Scenario 2	60.833	18.052	75.167	21.476	98.570	20.536
Scenario 3	75.944	25.504	61.500	10.656	100.164	16.762
Scenario 4	63.643	26.186	65.429	18.725	93.548	24.828
Scenario 5	56.500	17.546	67.857	20.852	90.785	17.250
Attempt Avg.	61.113	23.991	64.005	20.535	90.983	24.210

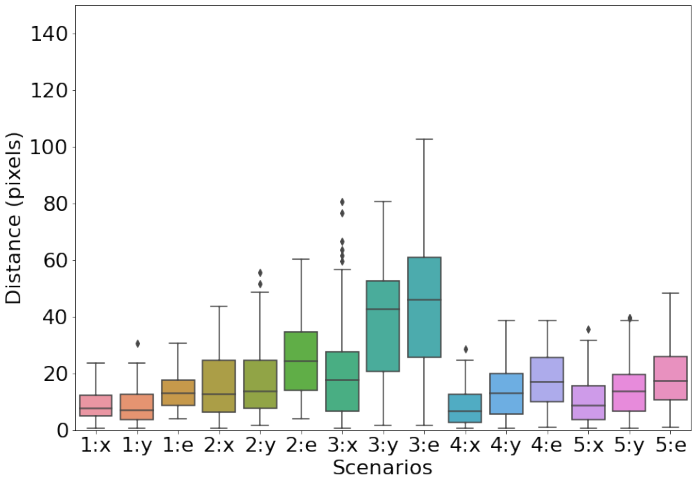
figures 4.3e and 4.3f. Furthermore, the tapping error variance and inter tap times variance per scenario for both test items are found in figures 4.3 and 4.4 respectively.

4.5 Tracing Accuracy

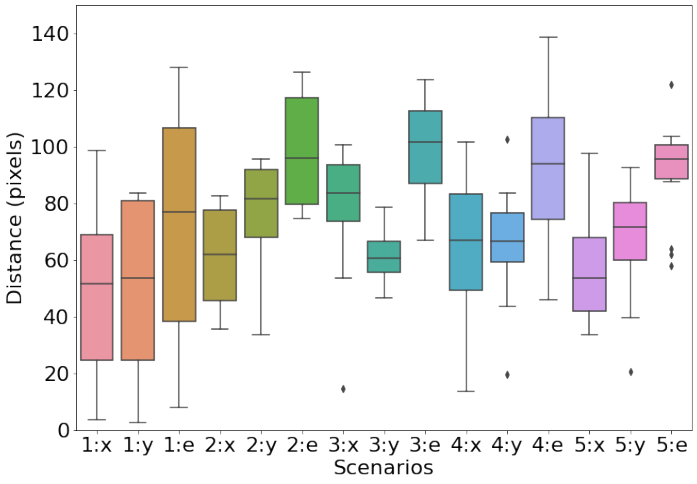
The test items in the tracing accuracy processing category are Connect The Dots and Tracing Line/Path. Figure 4.5 illustrates the distance error per segment, where a segment is a line drawn between two points (thus the distance error is the deviation from that line by the participant). Figure 4.6 contains the distance error variance per scenario for the Tracing Line/Path test item.

4.6 Image Analysis

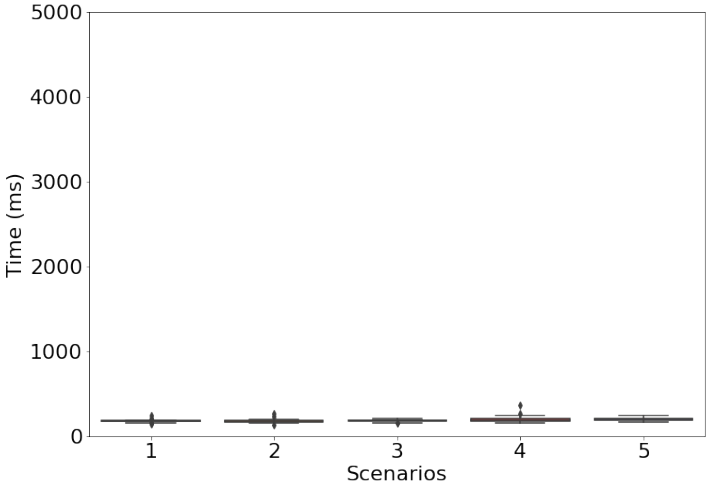
The image analysis processing category contained Colour Between Lines and Draw Object Given. For the Colour Between Lines test item, figures 4.7a and 4.7b illustrate the image as the participant coloured it in, and figures 4.7c and 4.7d indicate the error pixels shown as white pixels. Only one attempt was made for the Draw Image Given test item as the different scenarios (having different stock images) could be compared to one another as "good" and "bad" attempts. The premise of this is that an image similarity metric should indicate that a drawn image is more similar to the stock image counterpart than the



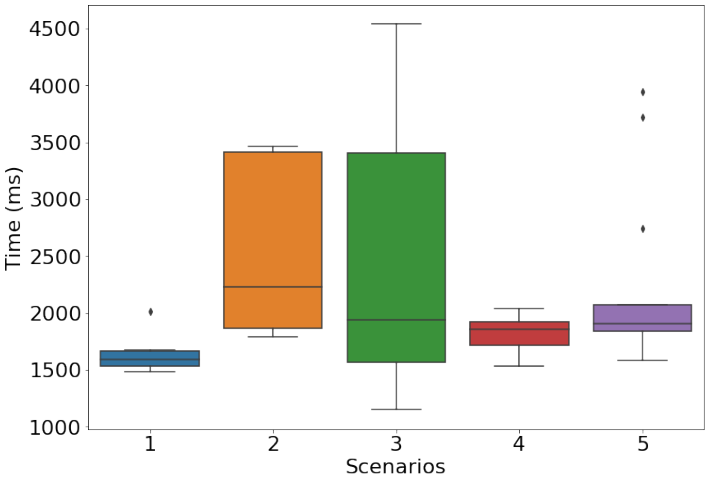
(a) Attempt 1 distance error variance.



(b) Attempt 2 distance error variance.

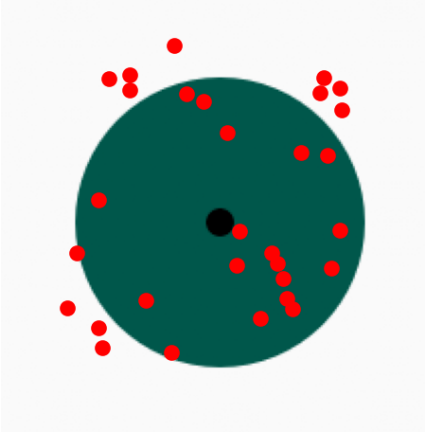


(c) Attempt 1 inter tap time variance.



(d) Attempt 2 inter tap time variance.

(e) Attempt 1 tap accuracy.



(f) Attempt 2 tap accuracy.

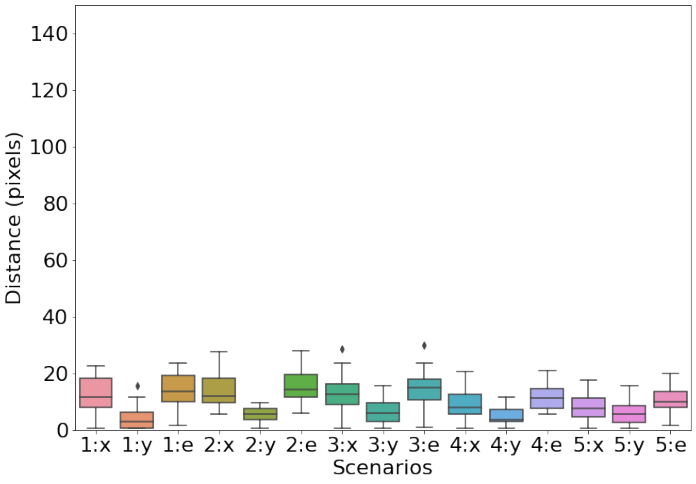
Figure 4.3: Graph results from the Timed Dot Tapping test item where attempt 1 is illustrated on the left and attempt 2 is illustrated on the right.

Table 4.5: Average tapping error per metric, per scenario for Rhythmic Dot Tapping attempts in pixels.

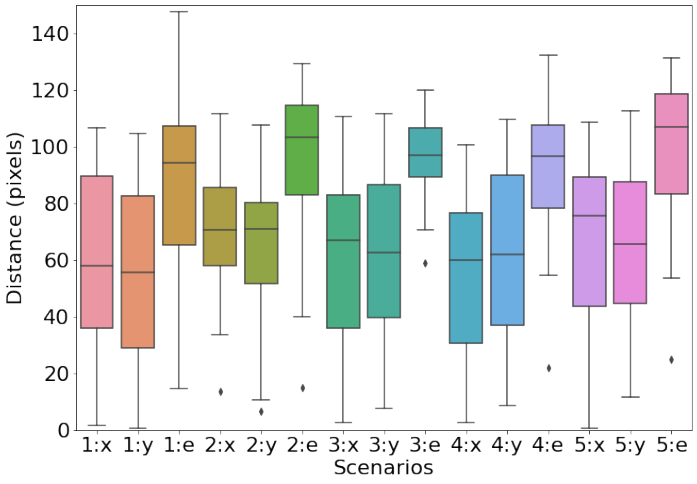
	Manhattan X		**Manhattan Y**		**Euclidean**	
	Mean	STD	Mean	STD	Mean	STD
Attempt 1						
Scenario 1	12.100	6.530	4.067	3.955	13.699	5.794
Scenario 2	13.933	6.070	5.400	2.413	15.243	5.799
Scenario 3	12.033	6.276	6.267	4.551	14.526	5.758
Scenario 4	9.200	5.392	4.900	2.984	11.197	4.610
Scenario 5	7.933	4.303	5.967	4.023	10.830	3.993
Attempt Avg.	11.040	5.714	5.320	3.585	13.099	5.191
Attempt 2						
Scenario 1	60.200	32.674	56.333	33.148	88.002	34.923
Scenario 2	67.667	21.613	65.833	26.380	96.847	26.391
Scenario 3	62.067	27.902	62.200	28.907	95.454	14.962
Scenario 4	56.600	29.231	62.767	30.825	91.877	22.503
Scenario 5	66.367	31.657	66.000	26.929	99.063	25.969
Attempt Avg.	62.580	28.615	62.627	29.238	94.249	24.950

Table 4.6: The Rhythmic Dot Tapping timing measures, average inter tap time per attempt, number of errors per attempt, and average time difference between stimulus and tap.

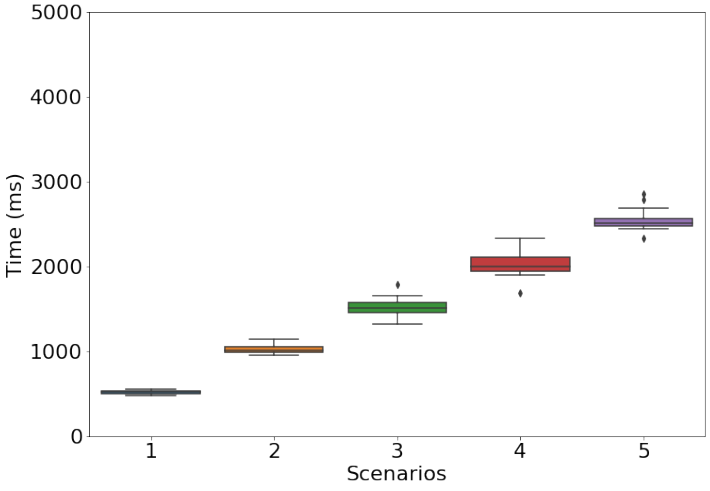
	Inter tap time		**Errors**	**Stimulus deviation**	
	Mean	STD		Mean	STD
Attempt 1					
Scenario 1	512.483	22.117	0	109.450	20.316
Scenario 2	1018.172	46.103	0	59.750	27.353
Scenario 3	1514.690	95.150	0	79.650	48.393
Scenario 4	2029.759	130.904	0	101.100	64.652
Scenario 5	2539.966	103.188	0	114.399	54.513
Attempt Avg.	1523.014	79.492	0.0	92.869	43.045
Attempt 2					
Scenario 1	513.552	283.433	1	141.200	83.401
Scenario 2	609.621	463.060	9	236.684	148.680
Scenario 3	539.069	507.568	13	429.600	182.054
Scenario 4	761.345	666.207	12	534.857	571.079
Scenario 5	559.828	414.098	17	594.222	388.090
Attempt Avg.	596.683	466.873	10.4	387.313	274.661



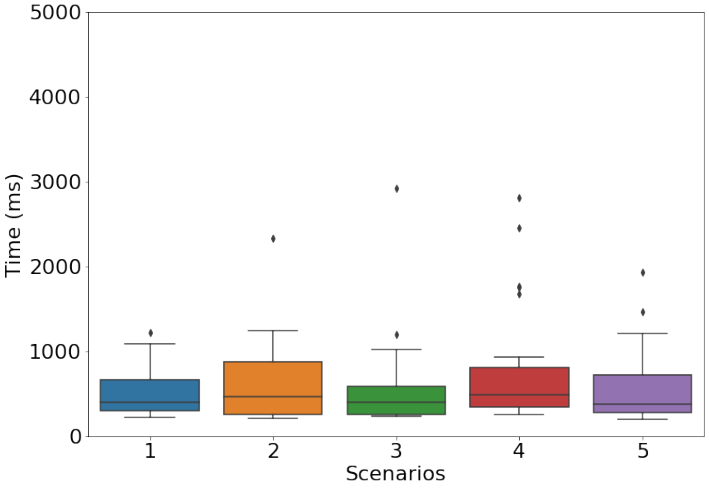
(a) Attempt 1 distance error variance.



(b) Attempt 2 distance error variance.



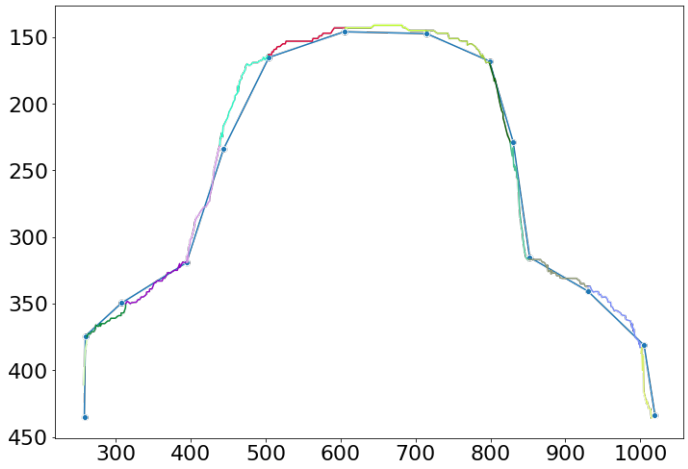
(c) Attempt 1 inter tap time variance.



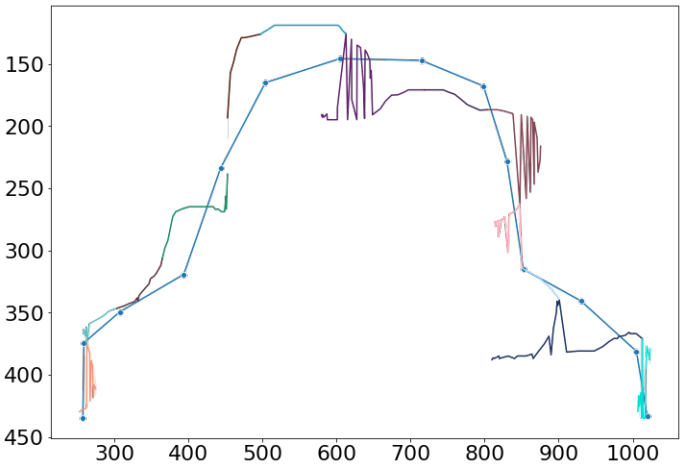
(d) Attempt 2 inter tap time variance.

Figure 4.4: Graph results from the Rhythmic Dot Tapping test item where attempt 1 is illustrated on the left and attempt 2 is illustrated on the right.

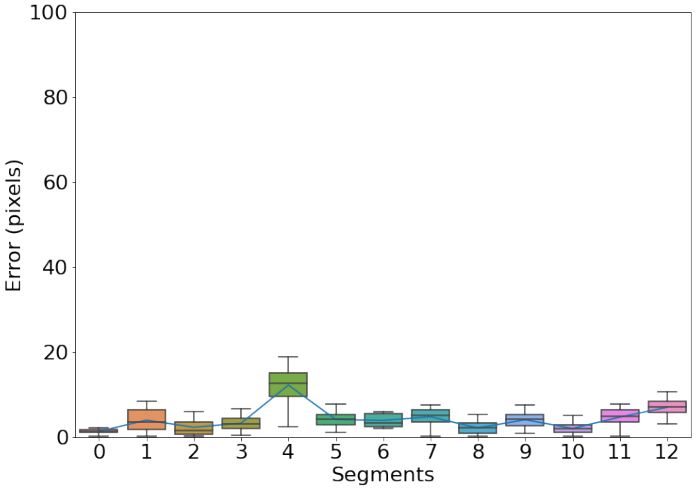
stock image is to other drawn images. Therefore the drawn circle can be seen as a "bad" attempt at drawing a triangle or cross, and vice versa. The results of the six measures used in the Draw Object Given test item are tabulated in table 4.7, along with the drawn and stock images illustrated in figure 4.8.



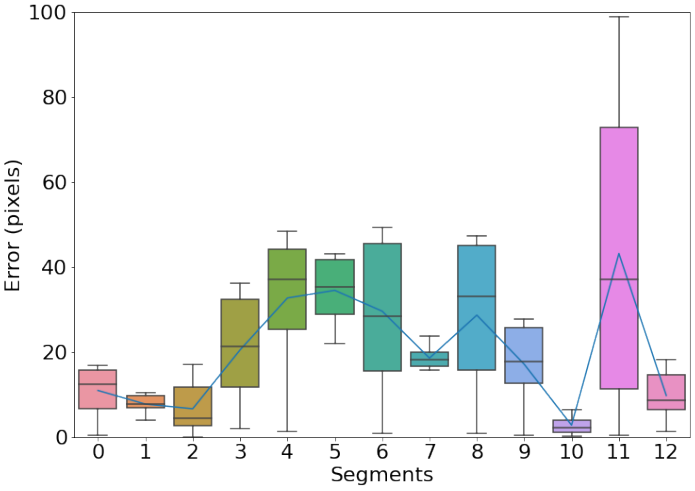
(a) Attempt 1 tracing illustration, with the tracing line coloured to indicate different sections belonging to different segments.



(b) Attempt 2 tracing illustration, with the tracing line coloured to indicate different sections belonging to different segments.

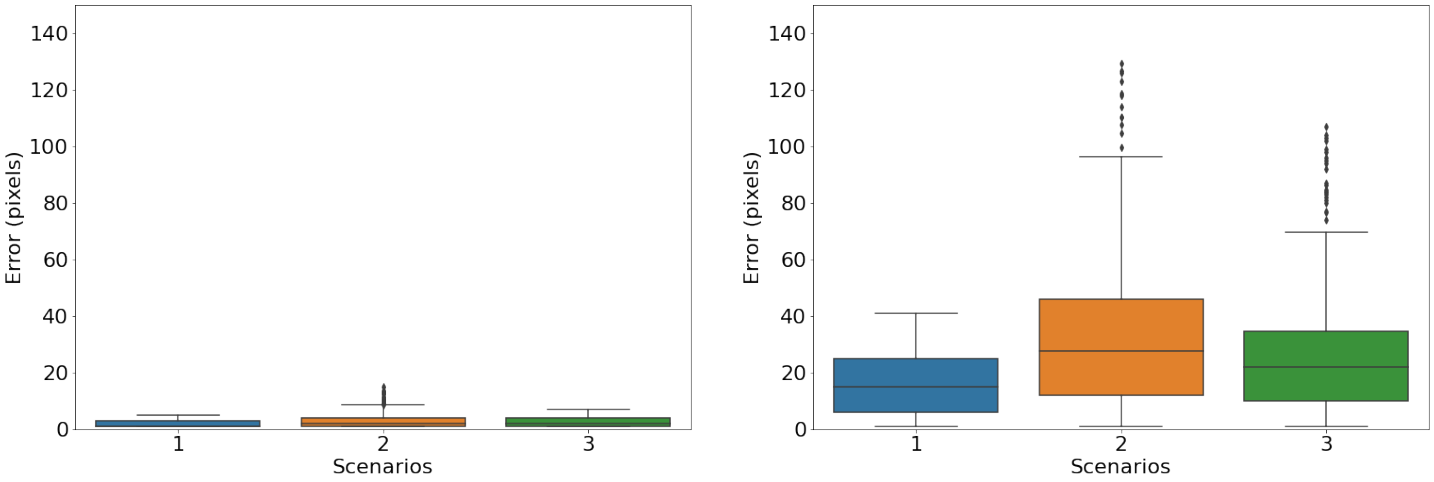


(c) Attempt 1 distance error variance.



(d) Attempt 2 distance error variance.

Figure 4.5: Error per segment plot for attempt 1 and 2 of the Connect the Dots test item, figure 4.5c and 4.5d, respectively. Attempt 1 has an average error of 4.221 pixels and attempt 2 had an average of 20.160 pixels.



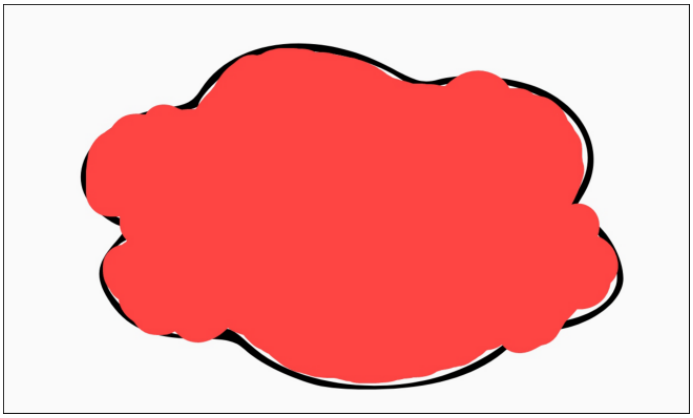
(a) Attempt 1 distance error variance.

(b) Attempt 2 distance error variance.

Figure 4.6: Error per segment plot for attempt 1 and 2 of the Tracing Line Path test item, figure 4.6a and 4.6b, respectively. Attempt 1 has an average error of 4.692 pixels and attempt 2 had an average of 50.891 pixels.

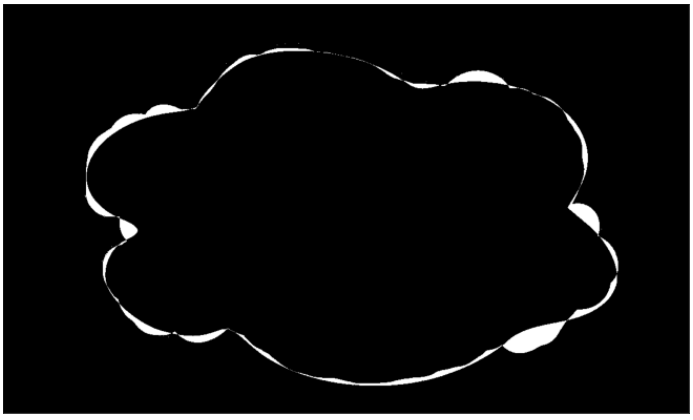
Table 4.7: Percentage similarity for each of the objects shown in figure 4.8. Each value was scaled and offset using the stock image compared to itself and the stock image compared to a blank image. Negative numbers indicate that the blank image is seen as more similar by certain metrics than the drawn image. **SSD** is sum of squared distances, **CS** is cosine similarity, **HD** is hausdorff distance, **SIFT** is scale invariant feature transform, **ResNet** refers to the modified ResNet-152 network, and **DeepAI** refers to the DeepAI image similarity API.

	SSD	**CS**	**HD**	**SIFT**	**ResNet**	**DeepAI**
Stock triangle compared to...						
Drawn Triangle	-249.825	-187.460	41.664	30.769	14.560	25.000
Drawn Circle	-301.191	-214.182	21.049	15.385	3.456	-6.250
Drawn Cross	-237.823	-179.436	32.998	38.462	4.878	-18.750
Stock circle compared to...						
Drawn Triangle	-237.589	-143.721	29.111	0.0	8.593	31.250
Drawn Circle	-291.813	-183.290	29.762	0.0	19.170	-6.250
Drawn Cross	-232.685	-175.478	19.211	0.0	2.455	25.000
Stock cross compared to...						
Drawn Triangle	-404.673	-1708.152	42.930	25.000	2.847	18.750
Drawn Circle	-485.065	-2575.406	21.042	20.000	-3.368	-18.750
Drawn Cross	-409.867	-1066.719	36.065	35.000	9.355	18.750

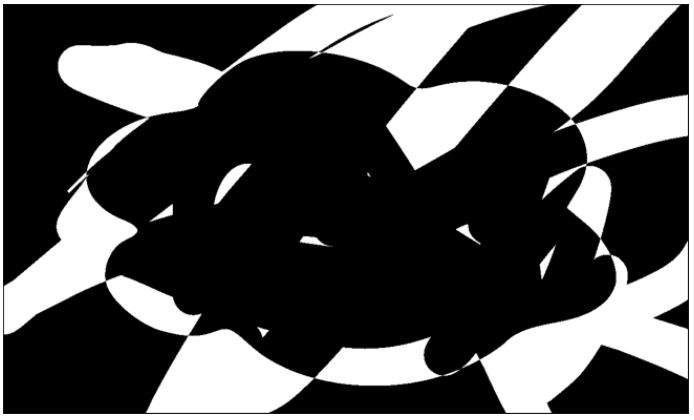


(a) Image captured from the Colour Between Lines attempt 1.

(b) Image captured from the Colour Between Lines attempt 2.



(c) Error pixel map for the Colour Between Lines attempt 1.



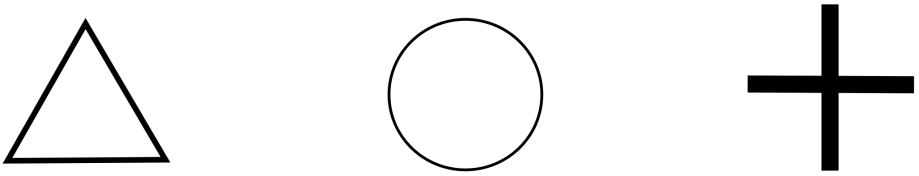
(d) Error pixel map for the Colour Between Lines attempt 2.

Figure 4.7: Colour Between Lines test item results. Figures 4.7a and 4.7b represent the images the used drew. Figures 4.7c and 4.7d represent the error map (after processing) with white pixels indicating error pixels. Attempt 1 (figures 4.7a and 4.7c) had a score of 97.731% and attempt 2 (figures 4.7b and 4.7d) had a score of 69.569%

4.7 Audio Analysis

The five test items in this processing category are Describe Picture, Give Opposite, Word Pronounce, Sentence Recall, and Number Recall. All five test items in this category of analysis require transcriptions from a speech-to-text model. The two speech-to-text models used were the pre-trained DeepSpeech 2 model and a Google Speech API. Of the two, only the Google Speech API transcriptions will be used as it faired better (more accurate transcriptions).

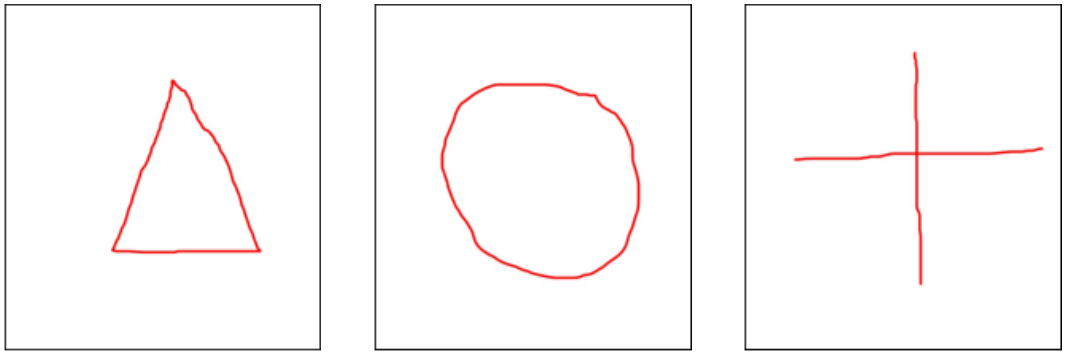
Two attempts for each of the test items were made, a "good" attempt and a "bad" one. The "good" attempt is meant to mimic a participant that can perform the test fully (such as repeat sentence word for word, remem-



(a) Stock image of a triangle.

(b) Stock image of a circle.

(c) Stock image of a cross.



(d) Drawn image of a triangle.

(e) Drawn image of a circle.

(f) Drawn image of a cross.

Figure 4.8: Draw Objects Given test item results. Figures 4.8d,4.8e, and 4.8f are images drawn in the application. Figures 4.8a,4.8b, and 4.8c are the stock images presented to the participant to be redrawn.

ber all number sequences, and explain image shown using all three keywords). This attempt was made by speaking naturally and knowing the correct sentence/number sequence/keywords to say. The second, or "bad" attempt, was meant to mimic a participant that is not able to repeat the sentence, number sequence, or word given, and also not able to give all keywords to describe the image. The desired result was acquired by saying the same as the previous attempt, but muffling the sound by covering up the researcher's mouth (the researcher performing the various attempts). This muffled speech resulted in the audio being severely distorted and the speech-to-text system not recognising the correct words.

Various transcriptions from the test items are included in table 4.8, illustrating the WER and CER metrics that were used. The Describe Picture test item in table 4.8 is the last entry, "a boy is sitting and reading". As the Describe Picture metric was calculated by checking for three keywords (in this case being "boy/man", "sit/sitting", "read/reading"), the score would be 33.3% for the second attempt, whereas the first attempt contains all three keywords.

Table 4.8: Word Error Rate (WER) and Character Error Rate (CER) of various recordings from the five test items. Each of the test items, besides Describe Picture, use WER and CER as metrics.

Attempt	Intended	Transcription	WER	CER
1	cat	cat	0.0%	0.0%
2		hat	100.0%	33.3%
1	house	house	0.0%	0.0%
2		blouse	100.0%	40.0%
1	I like dogs	I like dogs	0.0%	0.0%
2		I will take	63.6%	66.7%
1	the park closed in the evening	the park closed in the evening	0.0%	0.0%
2		sparklers in evening	46.7%	66.7%
1	A tree has leaves	A tree has leaves	0.0%	0.0%
2		but really	76.5%	100.0%
1	A boy is sitting and reading	A boy is sitting and reading	0.0%	0.0%
2		we're sitting in really	50.0%	83.3%

Chapter 5

Discussion

5.1 Interpretation of Results

The results listed in chapter 4 are only meant to serve as a verification that the tablet application and accompanying processing pipeline work, and to what degree they work as indented. Validation and reliability testing is an integral part of developmental assessments and is the final step after design and implementation. In order to validate a developmental assessment application, it needs to be administered to eligible participants and validity, and reliability measures need to be calculated. In light of the COVID situation of this year (2020), testing on preschool children was advised against. Therefore, more focus was given on the objective and autonomous analysis of the results.

Furthermore, the objects and words (any stimulus) were used merely as a placeholder to demonstrate the capabilities of the application and accompanying processing pipeline. Although the placeholder stimuli are selected and curated for assessment purposes, they might not suit further studies and their objectives. Simple images such as coloured shapes were selected as being age-appropriate (Howard and Melhuish, 2017), words used by speech therapists (De Beer, 2019) were selected, as well as images to be described. The stimuli presented can affect the results if the participant is unfamiliar with the stimulus or the stimulus contains any ambiguity. When constructing the test, there should be guarded against biasing results through the use of specific stimuli, using only cultural and age-appropriate stimuli.

All results are interpretable only as well as the participant understood what is expected of them. If the participant did not understand what is needed of them and performed the test item, the results would not yield an accurate representation of the construct being measured. The degree to which the participant understands the stimuli needs to be noted and taken into account when administering the test.

The premise of each test item's attempt ("good" and "bad") will be given along with how each result is interpretable.

Each of the test items has its raw results and no single score. The idea is to combine all the raw results of a test item into a coherent score able to indicate the participant's ability, while still giving researchers access to the raw results for further analysis. In order to construct such a single score, the role of each test item's raw results needs to be assessed and its ability to predict the participant's ability determined. Each of the results needs to be validated to ensure it is, in fact, predictive of the construct being tested, and that it can be combined, as is, with the other results to generate a singular score. That being said, the results are still interpretable to verify the workings of the application.

The first processing category in the results section is the Option Selection processing category. The category presents the participant with a stimulus and a set of options to select based on the stimulus. Table 4.1 shows the outcome of the scenario (whether or not the correct option was selected in the end), the number of options selected during the scenario, and the time it took to select the final answer. Whether or not the correct option was selected is the main result of these test items, which requires the participant to have understood the prompt, stimulus, and objective and selected the correct answer. It is also essential to determine whether or not the participant has selected multiple answers. Multiple answers can indicate possible confusion or hesitation that can warrant further investigation. The last metric tabulated is the time taken to select the final answer. The time taken to final answer indicates how long the participant takes to process the information and make a decision. Furthermore, the time to final selection can also be an indicator of the quality of the data. If the selection is made instantly, it might be an indicator that the participant has seen the test item before and has learned the response. Thus the result should not be used. Stimulus randomisation can be used to counteract the learning effect, but this can also be subverted. In order to protect against the learning effect, caution needs to be taken when assessing children as not to assess them with the same test multiple times and including enough variability in stimuli. One attempt was made per test item whereby "good" and "bad" performances were recorded. The first scenario mimicked a participant that is presented the stimulus, understands it, and quickly selects an answer. A quick correct answer can indicate the participant's receptive language skill, as faster processing speed with regards to language would allow the participant to perceive the stimulus and form an answer quicker (Leonard *et al.*, 2007). The second scenario mimics a participant that is still able to comprehend the stimulus but takes a long time to select an answer. Both scenario three and four mimic a participant that selects the wrong answer, also varying the time taken to make a decision. The fifth scenario illustrates how it would look if the participant selected many options during the scenario, either indicating hesitation or confusion.

The placement accuracy category envelops the Place Object Exactly and Build Object test items and pertains to the measurement of how accurately

objects were moved and rotated to fit the desired location and orientation. Both test items recorded the initial and final locations and orientations (X, Y coordinates) of pieces. The difference between the initial and final values are used to calculate the error using distance metrics, Manhattan and Euclidean, which are displayed in table 4.2 and 4.3. Table 4.3, which is the results table for the Build Object test item, also contains the number of pieces in the puzzle, which is necessary information when viewing the distance errors as more pieces can contribute to the error and should be taken into account. Lower distance errors for any of the two test items would correlate with better fine-motor skill as the participant was better able to place the object(s) in the correct or desired locations. More specifically, Build Object also gives insight into the participant's visuospatial intelligence. Visuospatial intelligence is used in order to determine how pieces moved around and placed should fit together in a larger picture, and is usually assessed by having the participant build objects with blocks or build puzzles. The more puzzle pieces present, the more spatial organisation needs to be done by the participant. Therefore the performance of the Build Object test item with increasing puzzle pieces can indicate visuospatial intelligence (which forms part of fine-motor ability) (Cameron *et al.*, 2012).

The dot tapping and time processing assessment category contained the Timed Dot Tapping and Rhythmic Dot Tapping test items. Both test items measured how accurately the participant tapped the dot, displayed in tables 4.4 and 4.5, and figures 4.3a, 4.3b, 4.4a, and 4.4b. The less the distances, the closer the tap was to the centre of the button. Thus the better the participant performed. Better performance would give an indication of better fine-motor skill, such as visuomotor integration (using one's visual perception to guide where the tap has to be). The timing component of these two test items differ. The Rhythmic Dot Tapping test item measures how well the participant can match and continue a rhythm presented. A person's rhythmic ability, in turn, indicates the participant's ability to estimate time (sub- and supra-second, depending on the configuration of the scenarios) and uphold motor timing rhythm (the rhythm of movement and timing the movements). Motor timing has a strong link to good overall motor performance (Falter and Noreika, 2011). The Timed Dot Tapping test item required the participant to tap as quickly as possible. Figures 4.3c and 4.3d indicate the variance of inter tap time of each scenario. Higher variance means less consistency in tapping accuracy when repeatedly tapping the button, which could indicate motor timing ability (Noreika *et al.*, 2013).

Tracing accuracy analysis consisted of measuring how accurately the participant could trace their finger on the desired line (or in the case of the Connect The Dots test item, between two given points). The analysis for the Connect The Dots test item was split into several segments, where a segment defines the area between two dots. In the particular Connect The Dots attempts shown in section 4.5 there were 14 dots to draw lines between and thus 13 segments.

Figures 4.5c and 4.5d indicate the variance and average of each of the segments graphically. Figures 4.6a and 4.6b illustrate the variance of tracing error by means of box graphs for each scenario of the Tracing Line/Path test item. Higher variance and overall distance error would indicate that the participant has traced the line or between two dots with less accuracy, thus having less fine-motor precision and control. Thus less distance error would correlate with a better fine-motor ability (Cohen *et al.*, 2018).

Image analysis pertained to the Colour Between Lines and Draw Object Given test items. For the former, the only metric used was the number of error pixels. The task given to the participant is to colour-in the object given on the screen without touching the borders or colouring outside the object. The error pixels indicates where the participant has wrongly coloured in the image. Colouring in an image requires fine-motor control (Wehrmann *et al.*, 2006); therefore, better fine-motor ability and control over fine movements would allow the participant to colour-in more accurately and lessen the number of error pixels. The Draw Objects Given test item aims to see how well the participant can copy by drawing the stock image given. Better fine-motor control would allow the participant to redraw the picture more accurately. Thus the similarity score between the stock image (presented as stimulus to the participant) and the drawn image can be used to indicate fine-motor ability (Vimercati *et al.*, 2015). The Draw Objects Given test item results seen in table 4.7 are the similarity scores of six different measures tested. The three drawn images were compared to each of the stock images (a cross, a circle, and a triangle), resulting in nine results. This comparison was made instead of having a "good", and "bad" attempt for each stock image as the circle can be seen as a "bad" attempt of the triangle and vice versa. In theory, the metrics should indicate that the drawn counterpart of each stock image is much more similar to its stock image than another stock image. Each of the scores was scaled and offset by the value of the stock image compared to itself and to a blank page, where the given image compared to itself would result in 100% and the stock image compared to the blank image would result in 0%. This scaling and offset explains the negative values seen in the table for SSD and CS measures. According to those measures, the blank image is more similar to the given image than the drawn image. Of all the metrics, the ResNet metric was the only one to consistently indicate that the drawn image of the corresponding image is more similar than that of unrelated images (the stock circle and drawn circle similarity score is higher than the stock circle and drawn cross/triangle). The SIFT algorithm does not perform well in the assessment of these images, which can in part be attributed to the lack of features. An image of the Eiffel tower would contain a lot more details and features that can be used to match it with other images. The images presented are simple, resulting in the SIFT method detecting only a handful of features to match. This decrease in features detected explains the zero scores for the stock circle and drawn images.

The processing of the Describe Picture test item requires checking the transcription acquired from the speech-to-text system for specific keywords. Each keyword present (out of a maximum of three) awards a point towards the test item's score. Each point indicates that the participant looked at the stimulus and understood an aspect of the image, which was then verbally communicated to the tablet, therefore indicating expressive language. The Give Opposite, Word Pronounce, Sentence Recall, and Number Recall test items are measured by the WER and CER metrics. WER and CER are measures of how accurate the transcription acquired from the speech-to-text model is compared to the desired text. Assuming the speech-to-text model is gives an accurate transcription (the case where the speech-to-text model does not give an accurate transcription is discussed later), each of the metrics (WER and CER) are interpretable as metrics of the participant's language skills (differing among the four test items). Table 4.8 lists each of the intended transcriptions (that which the participant has to repeat, or in the case of Describe Picture an entirely correct description of the picture), along with both attempts' transcriptions. The processing pipeline generates this table to give a quick, concise overview of each of the test item's scenarios. For the Give Opposite test item, lower WER and CER rates would indicate better vocabulary understanding (if the stimuli and content are culturally and age-appropriate), but also receptive language as the participant has to understand what is being said in order to process and verbally respond with an opposite (Viding *et al.*, 2004). The Word Pronounce test item would contain a lower CER for correct pronunciations of letters and have the WER to see whether or not the pronounced word is the desired one. The Sentence Recall test item would contain a lower CER for correct pronunciations of each word but also a lower WER and CER rate for the correct words in the sentence, indicating verbal working memory. Similar to Sentence Recall, the Number Recall test item has a lower CER and WER for correctly remembering the order and numbers that were to be reproduced verbally, also indicating verbal working memory (but with regard to numbers) as numbers and words are processed through different pathways (Carreiras *et al.*, 2015). These four test items can have ceiling and floor effects when the incorrect content and stimuli are used. Using words and stimuli that the participant has not seen before, or are unfamiliar with, would result in the participant being unable to complete the test items. Similarly, if the content used is all considered at the same difficulty level, and the participant can complete the test on that difficulty level, it would result in a ceiling effect.

5.2 Comparison to Previous Literature

Many of the test items within the tablet test in question are similar to test items in classical development assessment, as mentioned in section 3.2.2. The similarities and differences between the specific test items are also mentioned,

along with reasons for the changes. The results mentioned in the previous section are comparable to the metrics of test items from which they stem (for example the distance error of the Build Object test item is comparable with how well a participant can build a tower with blocks in the DDST).

When comparing tablet assessments and classical assessments, some positives and negatives arise for both assessment strategies. Concerning the assessment of motor skill, gross-motor and manipulation-based fine-motor are not as easily assessed with a tablet assessment as with classical assessment. Gross-motor is an integral part of motor assessment, where professionals administering the tests would gauge how well the participant's gross-motor ability is by having them perform tasks such as sitting down on the floor and standing up, catching a ball, and hopping on one leg. These assessments are not transferable to the current tablet assessment framework detailed within this thesis. Manipulation-based fine-motor test items entail the participant being asked to pick up a particular object and manipulate it before placing it down again. Manipulations may include rotating an object, threading beads onto a string, or putting coins in a jar. These assessments are difficult to transfer to a tablet test, and therefore manipulation-based fine-motor is not tested in the Build Object and Place Object Exactly test items. Visuospatial intelligence and fine-motor precision are still measured, but the manipulation aspect is lost.

Another problem with tablet assessments is that participants have to be familiar with a tablet. The need for familiarity is a problem in rural areas where the children have not been exposed to tablet technology, and the unfamiliarity can affect their performance on the tests (Howard and Okely, 2015). Finally, meta-analysis is not always possible in the context of the tablet testing framework. Meta-analysis refers to someone analysing the participant while they perform tasks, but not with regard to the task itself. An example meta-analysis is asking the participant to name objects, and while the participant is naming them, the administrator analyses how each letter is pronounced. Another is to observe where a participant's eyes gaze when performing specific tasks such as catching a ball. It is the analysis of how the participant interacts with the test itself. As one of the positives of the tablet assessment is that there is no need for a medical professional to facilitate the test, there will be no one able to perform meta-analysis.

There are several positives to tablet assessment. As just mentioned, the tablet assessment can be performed by non-trained medical professionals. This property of tablet tests broadens the reach of assessment to areas where medical professionals are not readily available, or the assessments are too expensive to afford. Distribution of the assessment is also not a problem. The application can be distributed to any device capable of running it, and active development can extend the number of devices by a large margin. The automatic analysis and generation of results, not the interpretation of the results, combats subjectivity that is present in classical assessments because the analysis pipeline

would assess the data the same each time. This consistent analysis is made possible by the use of objective measures such as distance errors and not subjective ones such as "how well the task was performed".

The tablet assessments mentioned in literature mainly focus on the ease of administration and objectivity. Although those are focus points for the specific tablet assessment in question, a difference is that this tablet assessment application is designed and constructed with modularity in mind. Modularity in this context means that objects that are shown and words read aloud can easily be swapped and changed for more age-appropriate or culturally appropriate objects. Cultural adaptations take time and effort (Nampijja *et al.*, 2010), and being able to switch out all resources used would ease this process. Modularity also means that the test item's behaviour can be altered slightly (for example, longer assessments or more options) in order to test other factors not foreseen by the current researchers.

The early years toolbox assessment by Howard and Melhuish (2017) contains an expressive vocabulary test item. This test item required the participant to name the object on the screen, still requires the person administering the test to assess whether or not the participant produced the correct label. Test items in the audio analysis category of the tablet assessment in question use a speech-to-text model to translate the audio and then score the measure, reducing subjectivity and bias. Furthermore, the DEEP tablet assessment by Bhavnani *et al.* (2019) measures taps (tapping a balloon five times, tapping alternating balloons, tapping any balloon on the screen) and the time of each tap, but leaves out coordinates of the taps, which is measured in current tablet assessment's Timed Dot Tapping test item. It also measures visuospatial intelligence with a two-piece puzzle snapping pieces into place, and not measuring exact error distance, such as with the Build Object and Place Object Exactly test item. Similar lack of in-depth assessment is seen in the tablet assessment by Pitchford and Outhwaite (2016), where the tapping tasks only measure the time taken to tap 30 times. Another test item requires the participant to build a pattern (measuring spatial intelligence) and scores the participant on whether or not the pattern is correct, not how the pattern was built (as measured in the Build Objects test item). Tablet devices can measure different metrics in parallel and are not constrained to measure only one or two metrics at a time. This property of tablet-based assessments needs to be harnessed, which can lead to richer and more plentiful data from developmental assessments.

5.3 Design Strengths and Weaknesses

The manner in which the test items for this tablet assessment were selected was from pre-existing classical, and tablet-based assessments for fine-motor and language selected according to multiple filtering processes and then im-

plemented. An alternative selection process for the test items is to map out all sub-domains of fine-motor and language and single out a test item for each of the sub-domains. This approach would ensure that the entire scope of both fine-motor and language is covered and ensure that no two test items measure the same construct. As previously mentioned, the application was built with modularity in mind. Scenarios and objects presented in the scenarios can be changed by changing the entries in a database, which is much less time consuming than altering the application itself.

Both Manhattan and Euclidean distances were used because Manhattan can highlight errors where the placement would be off centre in just one axis (for example, the participant always places the object just below the desired location). When a more condensed answer is required for a quick overview, the Euclidean distance can be used.

All of the Option Selection category test items assess receptive language, but each in a different way. Choose Associated Words gives the participant an object (which is a visual stimulus) and a set of words on buttons to select. Preschool children are most likely not able to read (as reading education starts in primary school), each of the buttons uses text-to-speech software to present the word to the participant verbally. The participant, therefore, has to match a visual stimulus to an auditory one. Choose Associated Object works in the exact opposite way, showing a word (and verbally presenting that word) and having the participant match it with a visual object. Visual and auditory stimuli start in different regions of the brain but are still processed by the language centre (Papanicolaou *et al.*, 1999), which indicates a clear need for the difference seen between Choose Associated Object and Choose Associated Word. Choose Picture works by giving a description of which object to select, instead of the name of the object. This further tests the participant's receptive skills by giving more than one criteria by which to select the correct answer (for example instructing the participant to select a blue square, out of a wide variety of coloured shapes) and resembles the "Not This" task implemented by Howard and Melhuish (2017). Object Recall tests non-verbal working memory, which forms part of general working memory and is also known as short term memory, as the participant is shown an image and has to recall it after a delay by selecting it from a grid of options. Verbal and non-verbal memory both have an effect on receptive language (Leonard *et al.*, 2007), and both need to be assessed. Follow Instructions further tests the participant's receptive language by determining if they can understand and follow instructions. By stringing multiple instructions together, the difficulty can be increased, and therefore testing receptive language more in-depth as difficulty following instructions is noted in receptive language impairment (Light *et al.*, 1998). The mechanism of measure is simple, the participant has instructions of what to do in the specific test item, a stimulus is shown to the participant, and an option is selected. The content of these option selection test items, however, is significant and should carefully be selected.

The Build Object and Place Object Exactly test items are handled by the same processing category but differ in that the former measures visuospatial intelligence as well as fine-motor precision. Both test items measure only the error distance of the objects from the desired location, which is affected by the participant's fine-motor skills. Furthermore, the test items that the Build Object and Place Object Exactly test items were based upon required the participant to manipulate objects physically. This physical manipulation entailed picking up objects, holding them in one's hand, rotating them, or performing other specified actions such as threading beads onto a string. This part of the assessment is lost when the test items were transferred to a tablet assessment.

The timed tapping tasks are a reliable estimation of manual processing speed and have been included in other tablet assessments (Pitchford and Outhwaite, 2016). However, the measurement of tapping accuracy on a stationary dot can be subject to the ceiling effect. The ceiling effect is not a problem but can be limiting in the measurement of the participant's fine-motor precision skill. Rhythm and time perception are a good predictor of fine-motor ability, and motor ability in general (Falter and Noreika, 2011; Noreika *et al.*, 2013), but might be hard for a participant on the younger side of the preschool scale to understand what is expected of them. Furthermore, timed tapping and rhythm-based measurements are being used more regularly in practical testing and assessment by occupational therapists (Rhode, 2019).

Albeit the two test items in the tracing accuracy processing category look different, they both test fine-motor precision, but there is a difference in how they measure it. The difference is visual feedback, the Connect The Dots test item gives visual feedback by showing the participant the line that is drawn, and the Tracing Line/Path test item does not. The difference is present to test the fine-motor precision with and without visual feedback as patients with deteriorating motor systems increasingly use visual feedback to compensate for the increased error in movement (Van Gemmert and Teulings, 2006). Furthermore, it was previously mentioned that preschool children are not able to read. As the path to draw in the Connect The Dots test item is numbered, it can result in difficulty for the child to follow the correct path. Therefore, in the current configuration of the test item, tasks will have to be explained to the participant.

The second to last processing category is the image analysis category containing the Colour Between Lines and Draw Object Given test items. The Colour Between Lines test item requires the participant to paint the inside of an outlined image by tracing their finger on the screen. Colouring-in an image enables the measurement of the entire range of fine-motor movements from very fine finger-movements (where small movements are necessary to colour-in, e.g. close to the image border) to larger finger movements and hand movements (e.g. colour-in the centre of the object). If combined with a stylus pen, this test item can be an excellent fine-motor assessment for school readiness

(Van Der Walt, 2019), as this would indicate the participant's ability to hold and use a pen. The use of a stylus pen is possible, and the current configuration of the application allows for it, but stylus pens are not readily available and do not work reliably on all devices. Moreover, a more intuitive way to indicate the area to be coloured in would be better as to avoid confusion. There is currently no indicator in the Colour Between Lines test item that shows where the participant must colour-in, and it is left up to the participant self or the person accompanying them. In order to ensure that the participant understands precisely needs to be done, an indicator needs to be implemented. Finally, the error pixels metric could be improved. Colouring-in a pixel on the border of the test item is a less severe error than colouring-in a pixel in one of the corners of the image. Weight should be attached to error pixels that increase or decrease their severity. The Draw Given Object test item processing metrics need to be improved. The modified ResNet approach was the only one that worked sufficiently well, indicating that the drawn version of a stock image is more similar to the stock image than the drawn versions of another stock images. Improvements can be made to have the system more accurately predict similarity scores.

The audio analysis of the five audio test items relies on a speech-to-text system transcribing the audio accurately. The dependency on a speech-to-text model needs to be taken into account when analysing the results generated from the processing pipeline. Although the results in table 4.8 were not tested on preschool children, a recent review (Gerosa *et al.*, 2009) highlights the current problems with speech-to-text systems concerning child speech. A child's speech is fundamentally different from that of adult people, and is commonly higher-pitched (Kent, 1976) and contains more inconsistencies (Gray *et al.*, 2014). Additionally, the amount of child speech data compared to that of adult speech data is far less (Claus *et al.*, 2013). It is, therefore, that speech-to-text systems have a lower recognition accuracy for children's speech when compared to that of adults. Lower recognition accuracy is further compounded by differences in pronunciations and accents for different groups of people. All these factors make it difficult to automatically test language ability in the way it is proposed in this tablet assessment. The speech-to-text system used in this project is by no means perfect and can influence the results based on accents and linguistic differences which results from one's setting. Further investigations into the effectiveness of the speech-to-text system will be done in future studies. Furthermore, both WER and CER metrics are prone to ceiling and floor effects. A better metric, or analysis system, would be to analyse phonemes. Phonemes are the core sound components of speech and together with other phonemes make up the sounds of letters and words. A distance metric could be used that measures the distance one phoneme's sound has from another. Therefore, the metric would become a distance metric of how far from the correct pronunciation words are. The need for this distance metric is further substantiated by the fact that what is said is not as important as

how it is said in test items like Word Pronounce.

5.4 Improvements and Future Work

Improvements to the modularity of the tablet application's set-up can be very beneficial. The current modularity lies in the database that is used to define the scenarios and the resource items the scenarios use. Along with the selection of test items in the Settings menu, a scenario editor would improve the modularity. The current way to change scenarios is to update the database (various methods can be used, but currently it is updated through the programming IDE Android Studio). Implementing a scenario changer would make the application much more user friendly and usable by researchers not familiar with Android development and database construction.

Placement accuracy test items can incorporate object manipulation-based testing. By instructing the participant to pick up the testing device and rotate it, move it, or place it in a particular position, this manipulation ability can be inferred. Mobile devices such as phones and tablets can use accelerometers and orientation sensors to determine whether or not the participant was successful in completing the task.

Similar to how the tablet's sensors can assess manipulation-based testing, meta-analysis can also be achieved. Using a mobile device's accelerometer and orientation sensor while the participant is being assessed might yield valuable information. Regarding eye fixation, the mobile device's front-facing camera can be used to determine where the participant's gaze is fixated, which might also yield valuable information.

In order to accurately assess whether or not participants can understand certain words, distractors can be used in the test items within the option selection category. Distractors are words that sound similar to the stimulus word, and objects that look like the stimulus object or are closely related to it. These would allow for more in-depth analysis of what the participant lacks in their receptive language ability. For example, to test whether or not the participant understands the word **tree**, the options could list a tree, a bee (sound distractor), a leaf (visual distractor), and an object not resembling or close to the sound of a tree, such as a rock. If the participant selects leaf, it will be recorded as being wrong but because of a visual distractor, and if the participant selects the bee, it will be recorded as wrong but because of an auditory distractor.

Measuring the distance error in the Timed Dot Tapping test item is a good indicator of how well the participant can use their fine-motor precision, but as previously mentioned, the button is stationary. In order to more in-depth assess the participant's fine-motor precision, the button can be relocated after each tap. This random relocation would force the participant first to look and assess where the button is before tapping each time.

The Connect The Dots test item should indicate visually to the participant which dot to draw to next as following numbered dots will be difficult for preschool children. As this task measures the participant's fine-motor ability and not comprehension of a larger image (having the participant see the unfinished image, understand what it represents, and then draw lines to complete it), or number recognition, the test item can indicate to the participant which dots they have to connect. The first dot could highlight at the beginning of the scenario, and as soon as the participant places their finger on it, the next dot illuminates. Illuminating the next dot to be tapped can continue until all dots have been connected.

Furthermore, the Colour Between Lines test item could clearly define where the participant has to colour-in by briefly changing its colour at the start. This indication would ensure that the participant can see where to colour-in and avoid confusion that would ultimately influence the test's results.

Further work needs to be done with regards to the Draw Object Given test item processing. The six measures presented were not sufficient with one measure barely indicating the desired results (that the drawn image of an object corresponds to the stock image of the same object). One suggested way is to build a dataset of drawn images and their stock image counterparts. This dataset can be used to either train a model like the ResNet-152 model from scratch or hone the pre-trained model in order to acquire better performance. This dataset can be further augmented by introducing translations (for example, shifting the drawn image to be off centre and to the left) and rotations into the data. Adding translations and rotations to the dataset will increase the robustness of an algorithm that would be able to detect whether or not the object drawn corresponds to the stock image. Additionally, the test items that the Draw Object Given test item was based on in classical assessment had an added complexity where they would show the stock image and then removed it, forcing the participant to draw from memory. This assessment dimension was not added but can be a further assessment.

In order to improve the assessment of the verbal language assessment test items, the speech-to-text system has to be improved. The way to improve it is by gathering data and training on that specific set of data. A dataset of pre-existing child speech corpora can be used, but will most likely not be from children in South Africa. For child speech data from South Africa, a dataset will have to be built from scratch. This data has to be within the context of South Africa because it has a wide variety of accents and languages that influence how certain words are pronounced.

There are numerous ways to validate a developmental assessment application, but the focus will be kept on how this specific assessment application and processing pipeline can be validated in future studies. Validation is crucial as it provides a backing that further research is valid and starts the processes of deploying the application for real-world use. First, a rigorous analysis will have to be done to determine what content is culturally and age-appropriate

in the context where this tablet will be assessed. Once the content is selected, several assessments will have to be done together with a gold standard classical assessment, which is suggested to be the Griffiths Mental Developmental Scales for its use in South Africa. Once the results of both tests have been acquired, correlation testing can be done in order to single out the test items and scenarios that mirror the results of the gold standard test's results. Only after that can any validity be attached to this tablet assessment.

Chapter 6

Conclusion

Numerous factors can affect a child's development, such as poverty, malnutrition, and lack of correct stimulation. The presence of neurodivergence and disabilities can further increase developmental delays. In order to help children cope with or mitigate these developmental delays, intervention programs are required. However, before intervention can be done, the problem needs to be known. Awareness of these developmental delays can be acquired through developmental assessments done by medical professionals, which are known as classical developmental assessments. These assessments are often expensive, time- and resource-intensive, and may have a lack of people able to administer them. These assessments are also susceptible to bias and subjectivity, as often the administrator has to gauge how well the participant is performing an action.

Tablet assessments counter the subjectivity by measuring non-subjective metrics such as distance and time. Tablet based assessment can reduce the cost as there is no need for a medical professional to be present when administering the assessment. These assessments also increase availability as they can be administering with a wide range of mobile devices.

A series of test items for both fine-motor and language were gathered, filtered, and adapted for the context of tablet assessments. Eighteen test items were identified, ten language test items and eight fine-motor test items. An Android-based application was built to house these test items. The application was built to be modular, which enables cultural and age adaptations do be made without reconstructing the entire application, thus speeding up the process. This tablet assessment allows for more in-depth assessment as multiple metrics are measured per test item. An accompanying assessment pipeline was constructed that can process the data from the tablet assessment into meaningful results for further interpretation. This pipeline removes the need for a medical professional to compile and process the results, making it more accessible to regions where medical professionals are not readily available.

Once the tablet application and processing pipeline have been validated against a gold standard assessment, the widespread use thereof can begin. The

tablet application can be distributed to places where classical developmental assessments are not practical because of cost and availability. More children can be assessed with regards to their fine-motor and language abilities, which will help identify problem areas and help allied health professionals plan and curate intervention programs. Ultimately, with continued development and research into tablet applications such as this one, children who suffer from developmental delays can more reliably be identified and receive the help they deserve.

Appendices

Appendix A

Tablet Assessment Additional Information

A.1 Test Item Images

This section contains example images of each test item as found in the final development phase of the tablet application. These images are used as a visual aid to help illustrate how each test item looks.

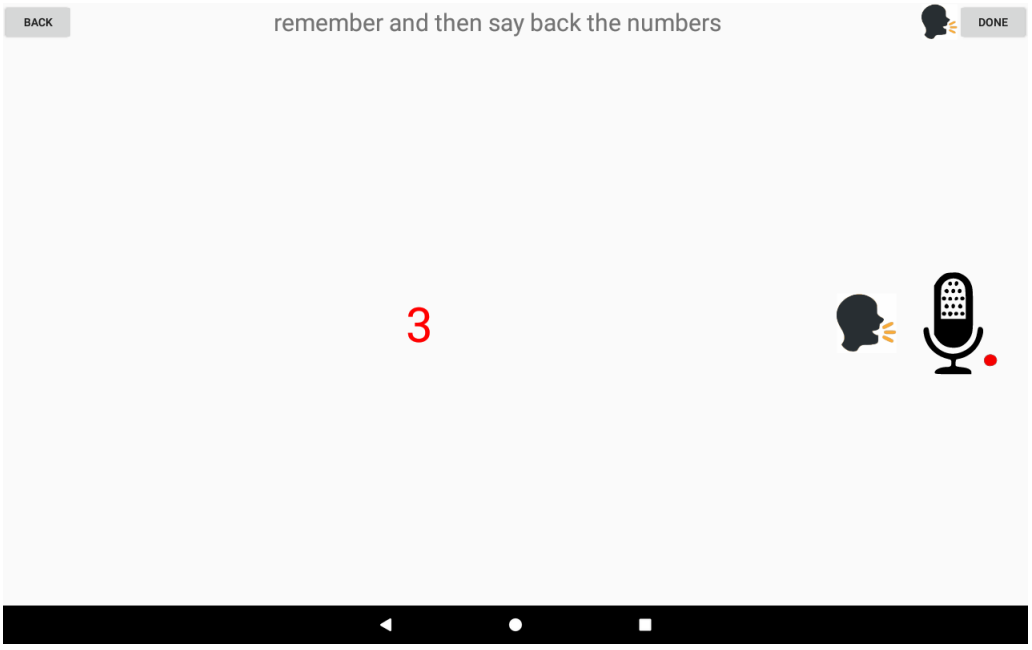


Figure A.1: Number Recall test item

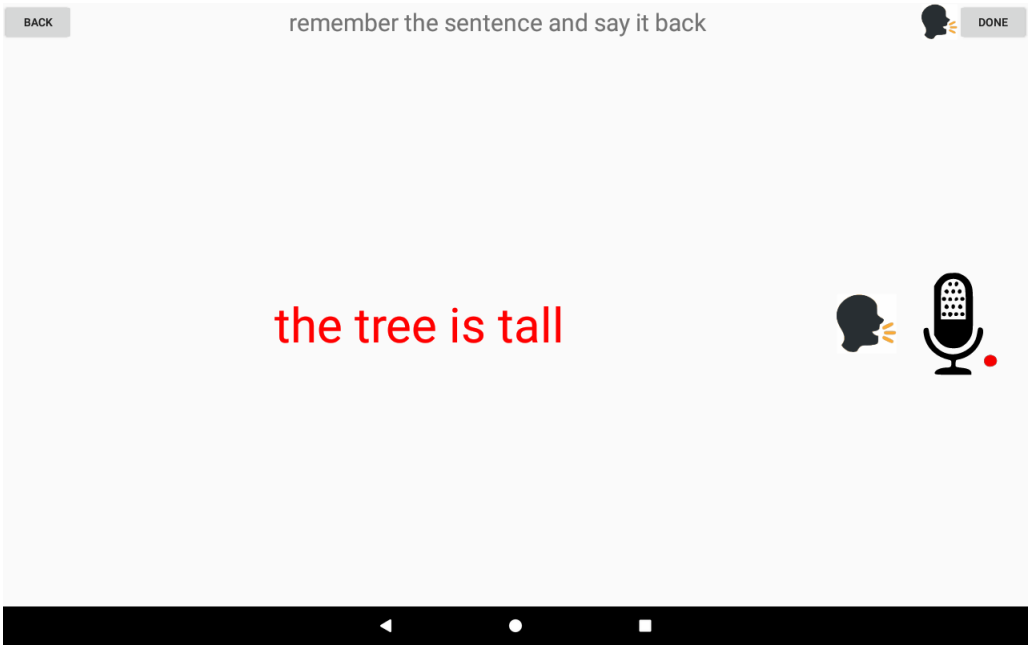


Figure A.2: Sentence Recall test item

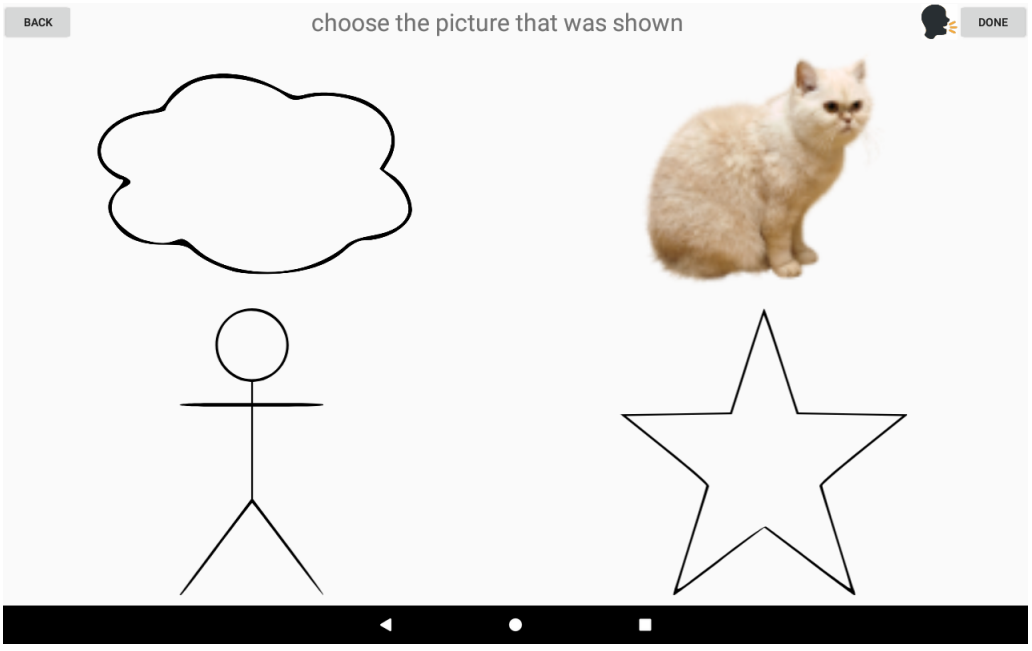


Figure A.3: Object Recall test item

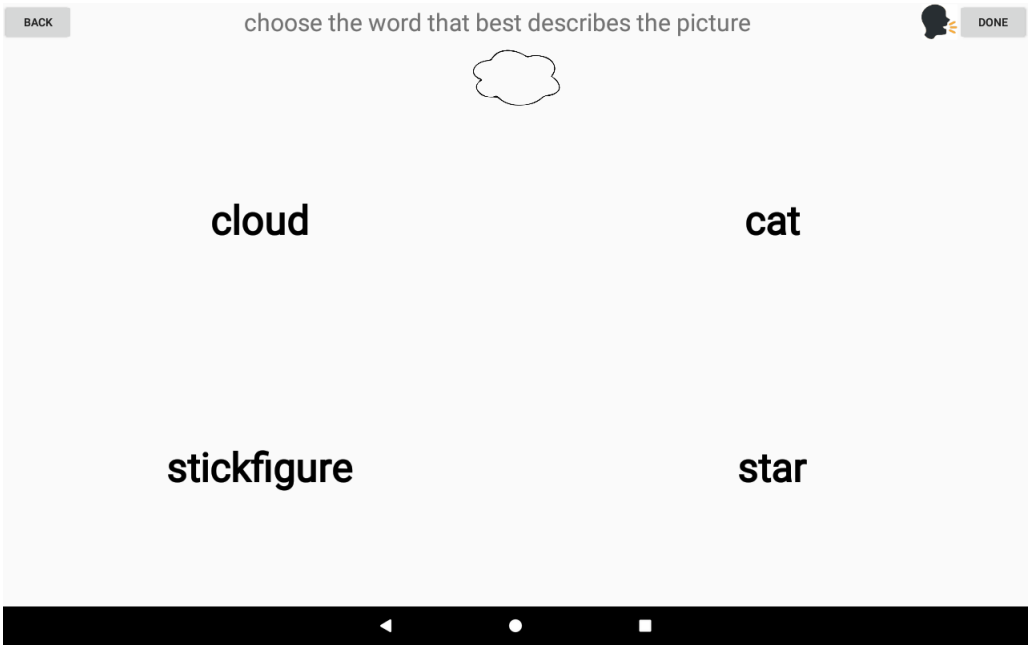


Figure A.4: Choose Associated Word test item

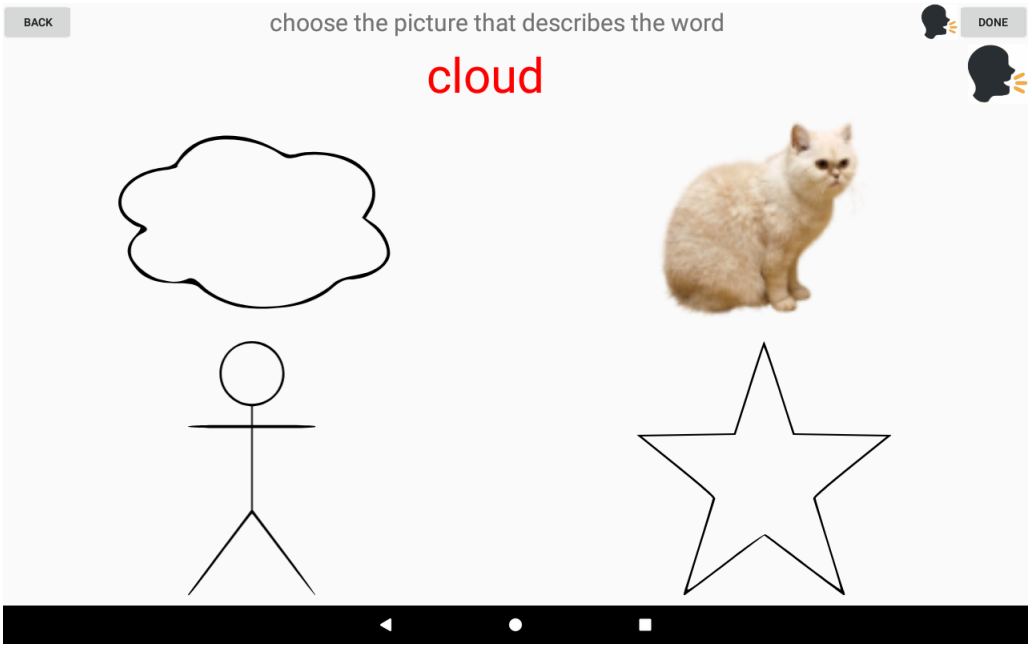


Figure A.5: Choose Associated Object test item

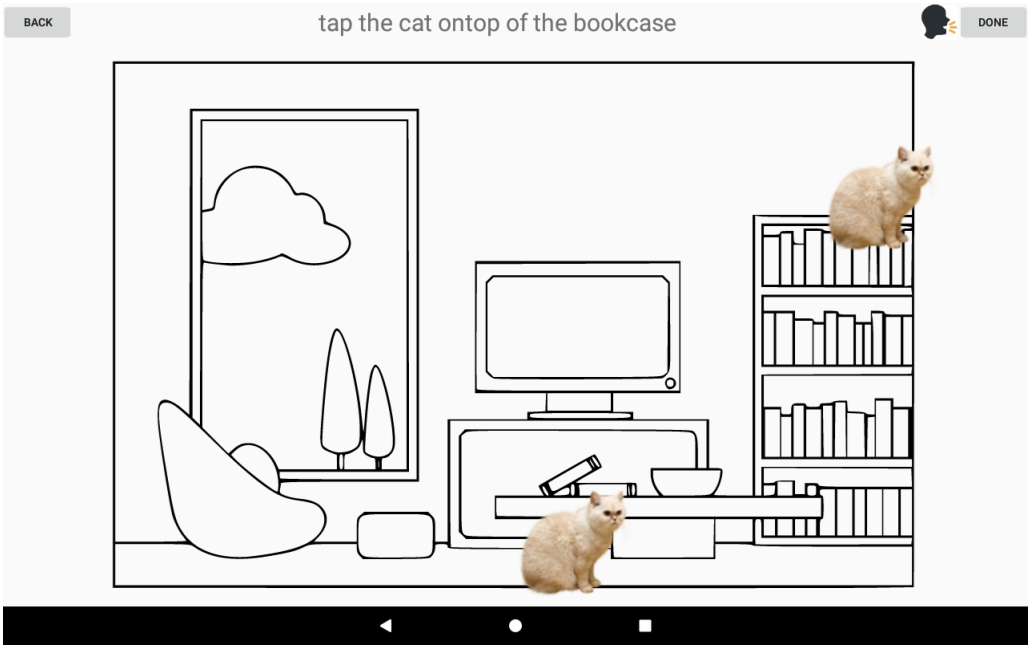


Figure A.6: Follow Instructions test item

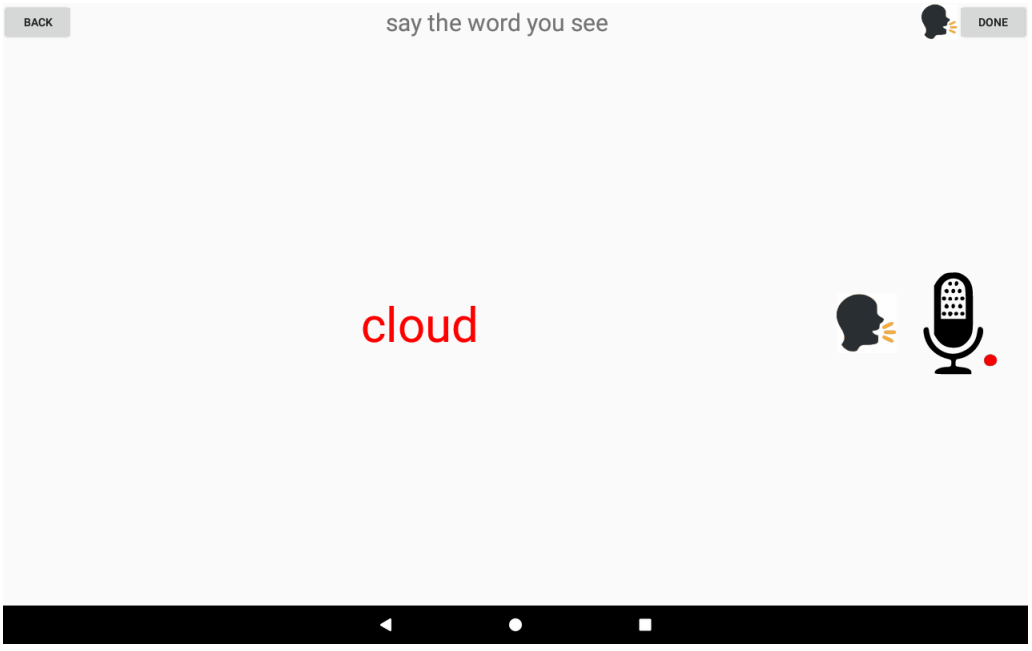


Figure A.7: Word Pronounce test item

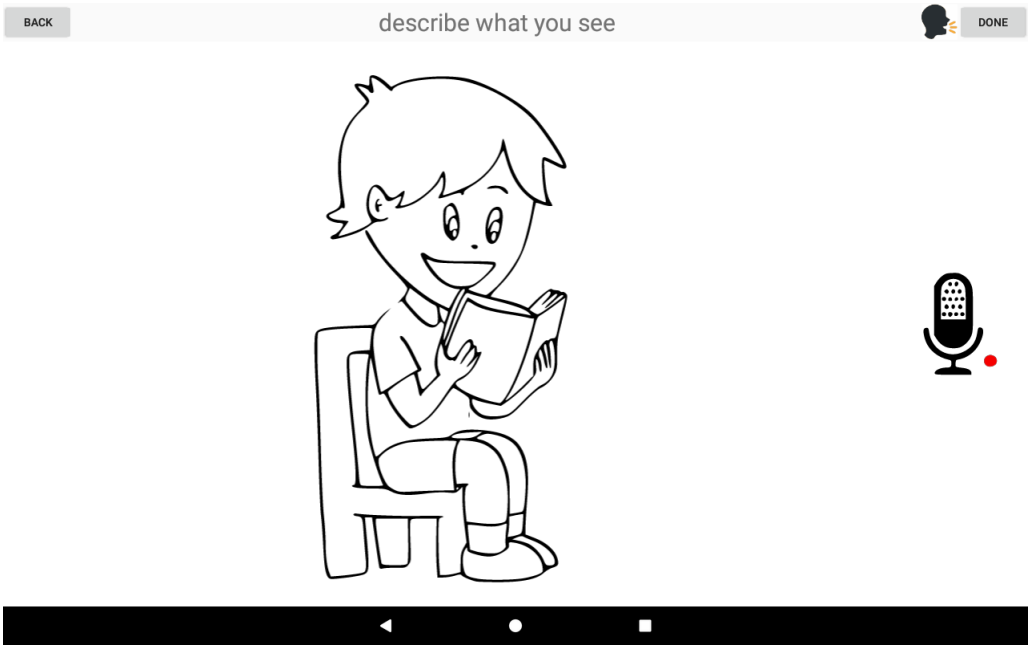


Figure A.8: Describe Picture test item

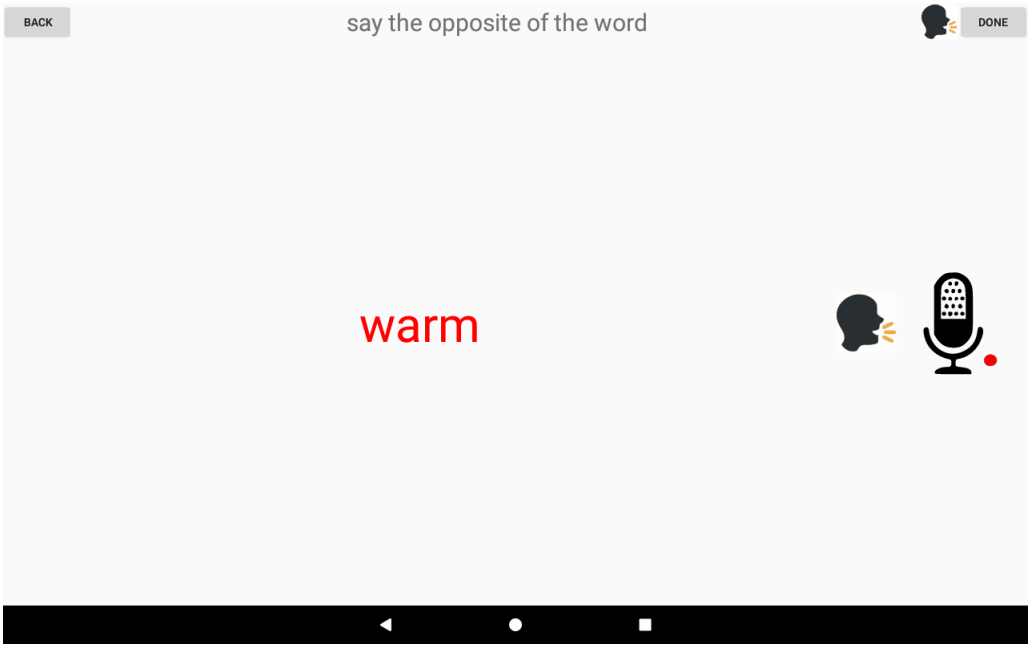


Figure A.9: Give Opposite test item

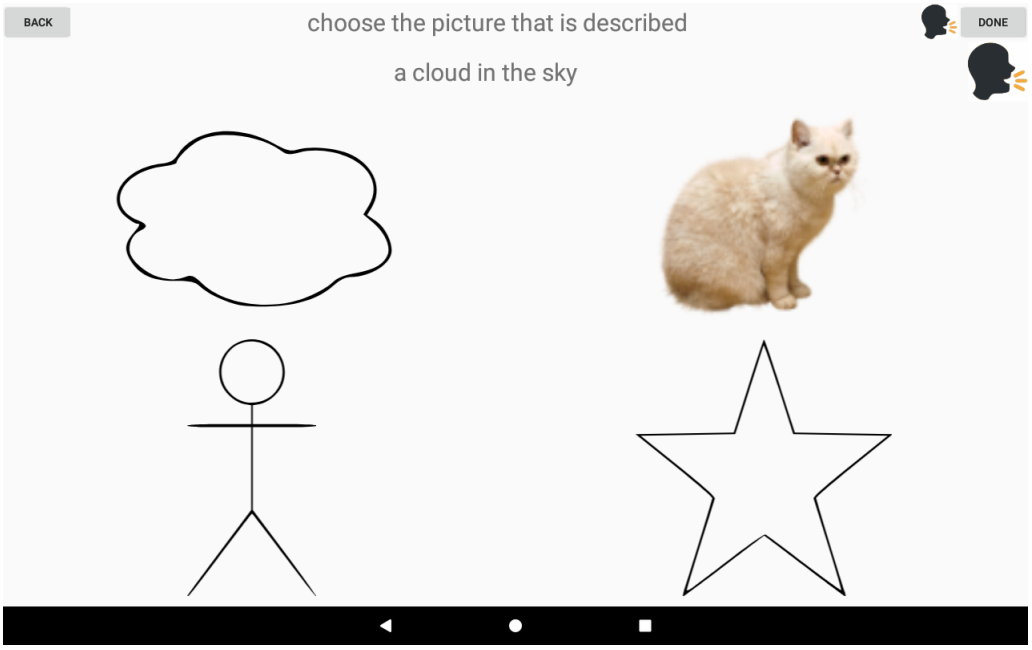


Figure A.10: Choose Picture test item

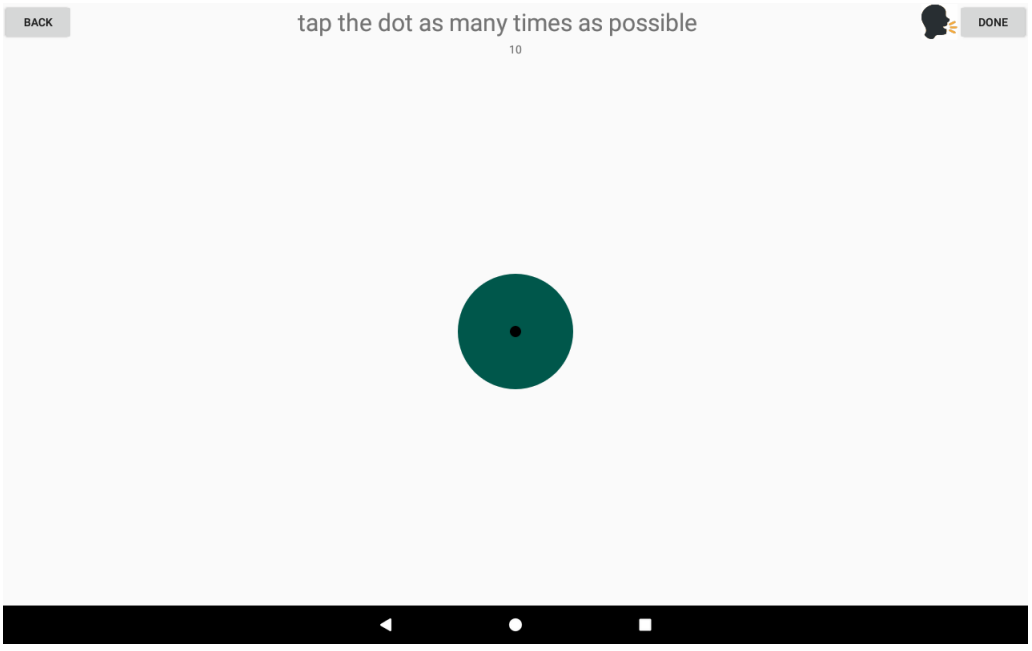


Figure A.11: Timed Dot Tapping test item

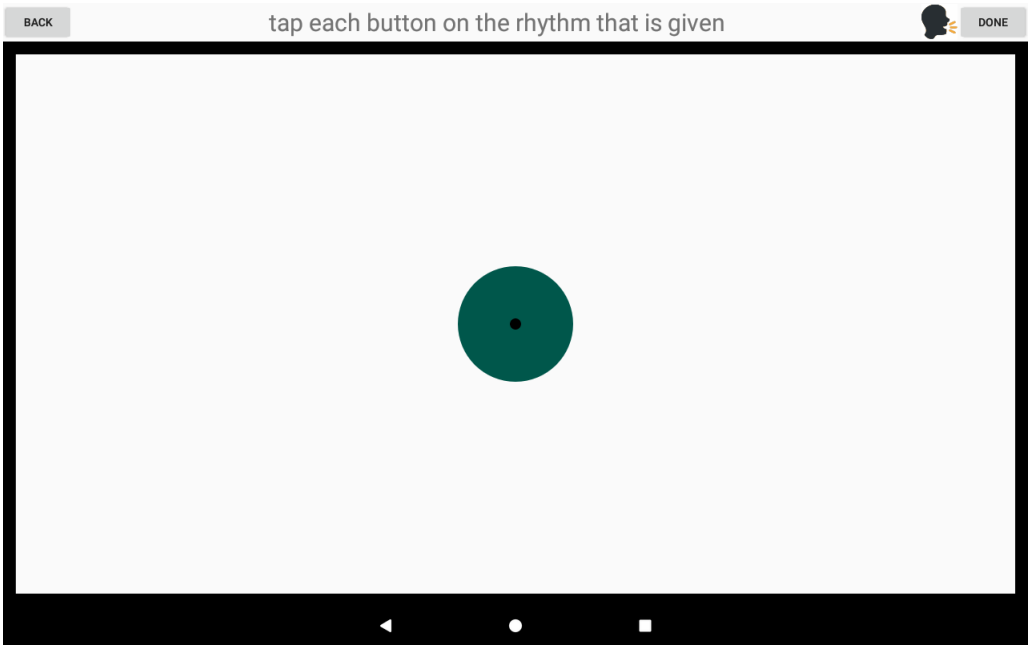


Figure A.12: Rhythmic Dot Tapping test item

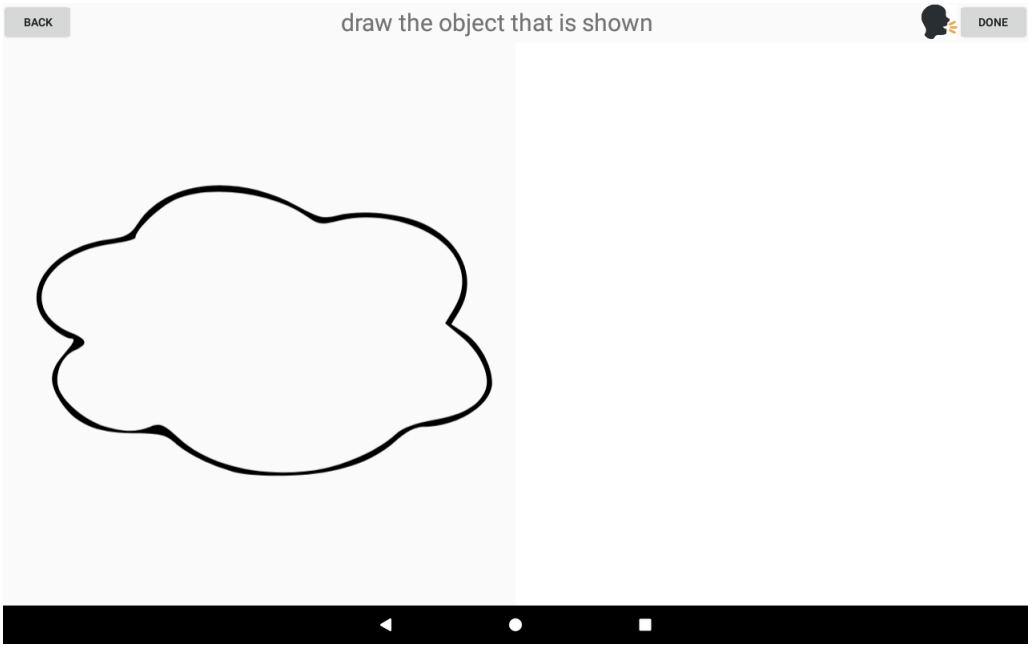


Figure A.13: Draw Object Given test item

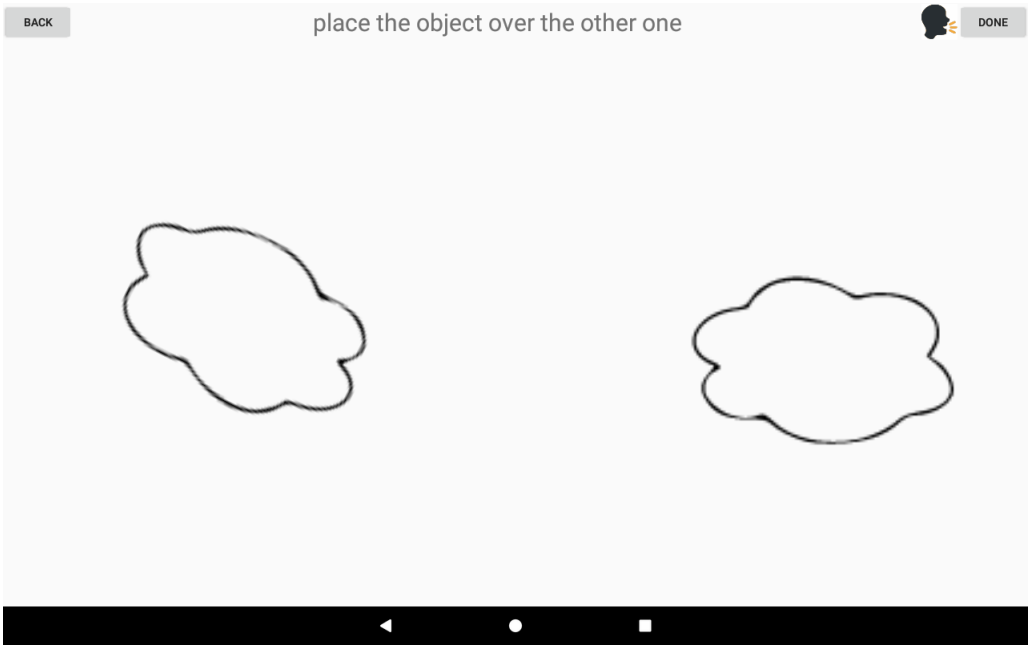


Figure A.14: Place Object Exactly test item

Figure A.15: Building Object test item

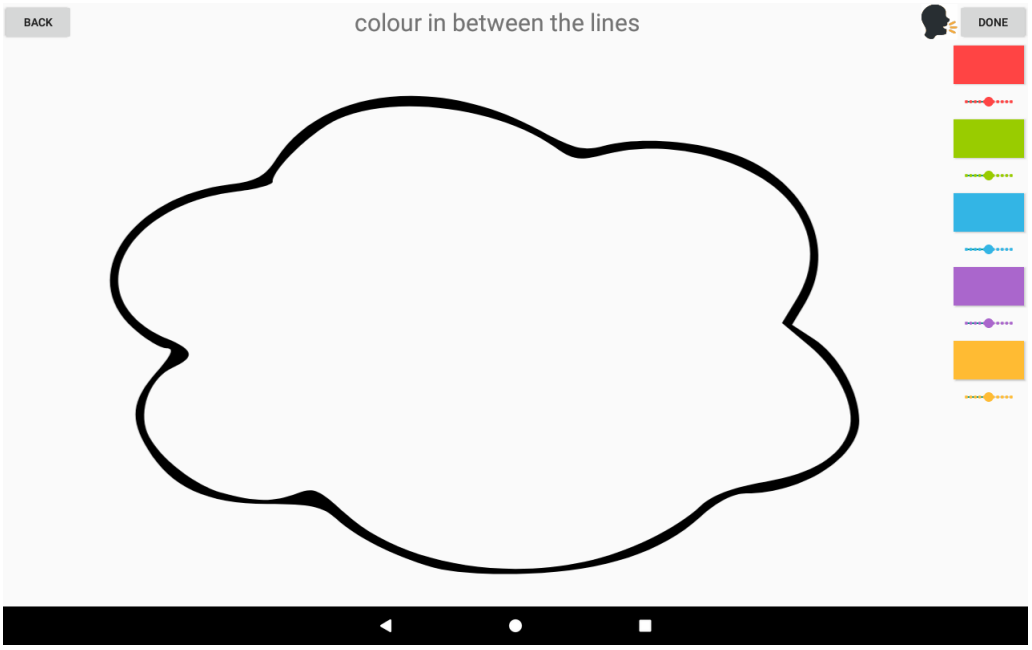


Figure A.16: Colour Between Lines test item

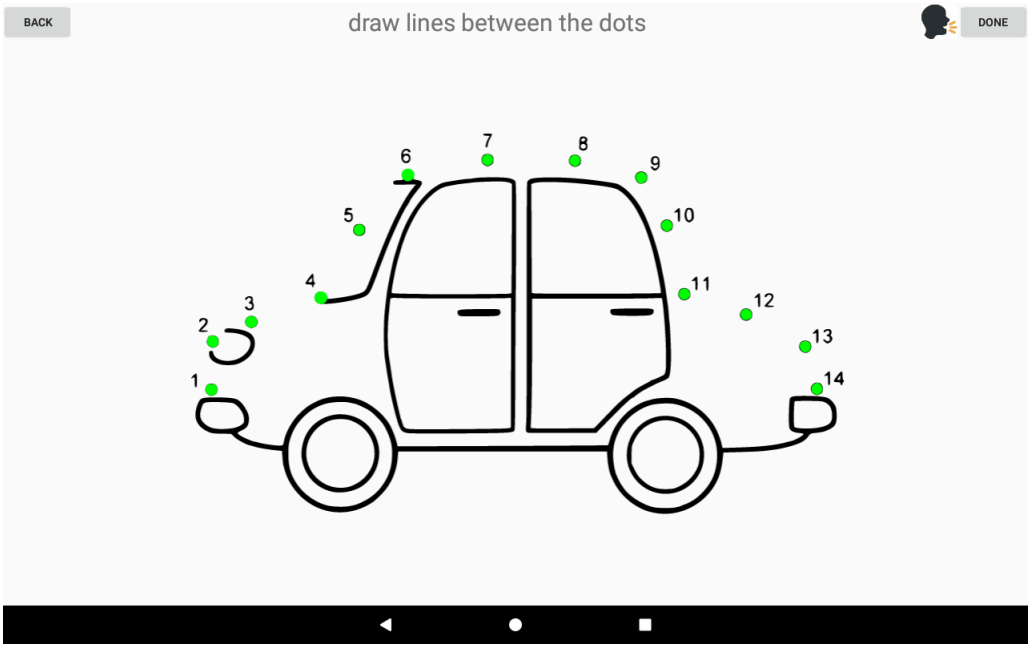


Figure A.17: Connect The Dots test item

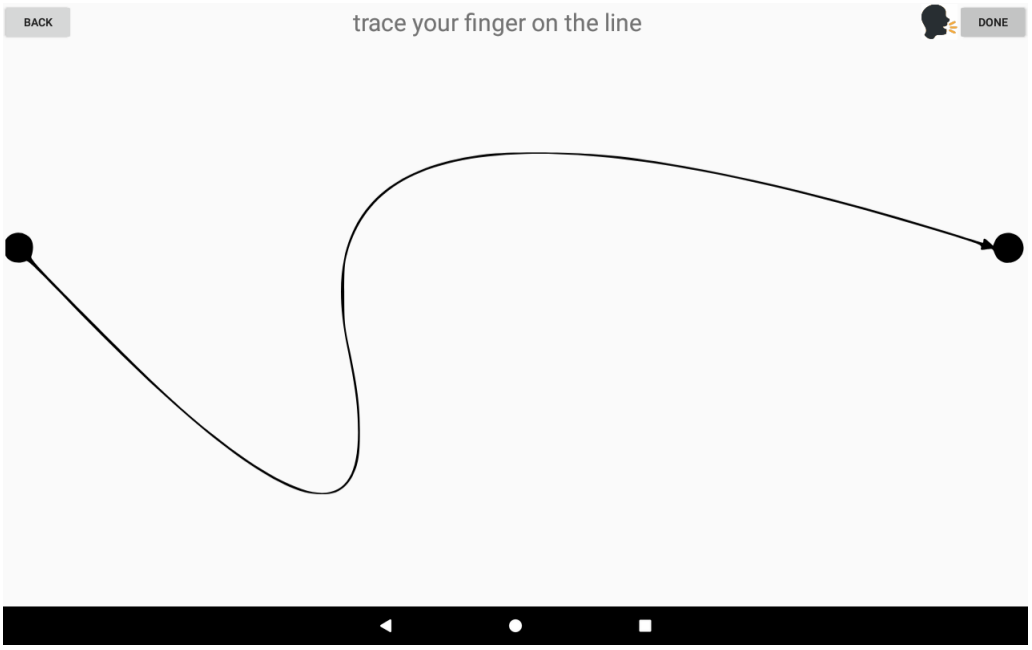


Figure A.18: Tracing Line/Path test item

A.2 Custom Components

Custom components are created by building on top of already existing components made available by the Android developer IDE. These components were customised to more easily monitor and track interactions on the tablet test. Hereafter is a list of all the custom components created to facilitate the construction of the tablet assessment in question.

AutoTable - The AutoTable receives as input the number of rows and columns it needs to generate, as well as what it should use to fill the grid. Used to display a grid of buttons, images, or words.

DescriptionTextView - Test items that require a description of an object to be displayed use this component. It receives as input a description text and audibly voices the description if tapped.

InstructionTextView - Present in the Activity hosting the test items, this component receives as input the instructions for a given test item and audibly voices the instructions if tapped.

MoveImageView - Receives a resource item and displays an object. It can be dragged, placed, and rotated on the screen and logs every touch, drag, and rotate movement.

ObjectImageView - Used to display a resource item's image and if tapped audibly voices the name of the object.

OptionButton - Used to display a word or an image from a resource item with the added functionality of a button such as highlighting when selected. It records each tap action performed on it.

PaintView - Used to record finger paths drawn by the participant, it is also able to display said finger path as if the participant had painted it.

RecordButton - Used in audio recording tasks, this component receives a recording resource item (typically showing a microphone as an image) and shows the recording status by having a red glow when it is recording. This component also stores the audio file and logs the location of it for later retrieval.

TouchButton - Used in the tapping tasks, it receives a resource item to display the button to be tapped and logs each tap location and time.

WordTextView - Used to display words and voice either the word itself or a description of the concept represented by the word. It logs each tap.

A.3 Broadcasts

Broadcasting is a way for components to send data to one another. Each component mentioned above can broadcast information to others. This application makes use of broadcasting to log data. Every interaction the participant has with the application is logged using a broadcast. Broadcasts use tags that can be listened for to indicate the contents of the broadcast. The broadcast identifiers below are used to sort and correctly save the data containing within the broadcast.

image_button_click - Used by any image button when clicked and it sends the following data: what resource item did the image button have when clicked, what was its unique ID, when it was clicked, where exactly it was clicked.

textview_click - Used by InstructionTextView, WordTextView, and DescriptionTextView, this broadcast sends exactly where and when a TextView component was clicked.

activity_start_time - As soon as the test battery starts, the Activity hosting all the test item fragments logs its unique ID and the date and time of the start of the test battery.

fragement_start_time - Each test item logs the time it is rendered, signalling the test item has started.

timer_activate - Some test items require a timer in the background to periodically perform a task. This broadcast is sent to indicate the status of the timer and to log the status within each test item's scenario events log. The timer signals when it starts, when it performs an action (known as a tick), and when the timer stops.

stimulus_firing - Used for only one test item, Rhythmic Dot Tapping, it sends this broadcast when the stimulus triggers and is displayed to the participant.

image_view_click - Referring to both MoveImageView and ObjectImageView, this broadcast is sent when a click action is performed on the aforementioned components. A click action is the action of tapping or placing one's finger on the button and promptly removing it. This action logs the number of times the component has been clicked, and where it has been clicked.

image_view_touch - Also referring to both MoveImageView and ObjectImageView, this broadcast is sent when a touch action is performed on the components as mentioned earlier. A touch action is the action of placing one's finger on the component for an extended amount of time, moving it around (where each movement is logged with this broadcast) and rotating it. This action logs the location of the touch event on the component, its location, its rotation, and what time the action occurred.

finger_path - Used with test items that require the participant to trace or draw a line/path, this broadcast contains the finger path data acquired from a PaintView component. Finger path data is a per time step (how quickly the tablet can record) location array of where the participant drew a line/path or traced their finger. Each of the location points contains a time as well.

resource_item - Sent containing the data of a resource item to log the use of said resource item in a specific scenario of a test item.

scenario_info - As each scenario is loaded from the database, it gets logged to appear in the scenario log of each test item. The scenario is logged along with a unique identifier and a time to differentiate better which scenario happened when.

recording_location - Used by the RecordButton component, it is broadcast when a recording has finished and is saved. The saved location is sent along to be able to recover the audio file for later analysis.

component_size_location - Each component that has a visual element sends this broadcast when it is rendered. It logs the location on the

screen when rendered, the width and height of the component, and its unique ID.

puzzle_piece_location - The BuildObject test item broadcasts this message once it receives an image and has divided it into the predefined amount of pieces. Each piece's location, or the location of the top left corner of the piece, is logged.

image_location - Used by test items that store images for later analysis, this message is broadcast when an image is captured from a test item and saved on the tablet. The file location is attached to this broadcast to be able to recover the image.

store_data - Broadcast by the hosting activity, this is sent when all test items are finished. This broadcast signals to the logging service that the test battery is finished and the JSON file must be saved, along with the time and date of the test.

A.4 Resource Groups

Resource items are sorted into resource groups to enable random selection of the appropriate resources. Listed here are the resource groups to which resource items can belong.

1) **normal objects** - These resources have words, descriptions, images of common household or known objects.

2) **describe objects** - More than a simple object, it is a person/animal performing an action.

3) **number objects** - Numbers from 0 - 9.

4) **opposite objects** - Opposites are given together separated by a unique character, such as "warm_cold".

5) **sentence objects** - Full sentences and identifying data for said sentence, with no images.

6) **puzzle objects** - Images for the use of building puzzles.

7) **dot objects** - Dot image to be displayed for dot tapping task.

8) **connect the dots objects** - Connect the dots images where each dot's coordinates are listed in the object's description.

9) **line or path objects** - Line and path images.

a) **background objects** - Images to be used as backgrounds along with locations to place objects on the background.

Appendix B

Artificial Neural Networks Overview

B.1 Introduction

This appendix is by no means a complete explanation of artificial neural networks and its derivatives as it is a broad field of study with many intricacies. It is indented to give the reader some basic knowledge and understanding of artificial neural networks, enough to understand how and why they are used.

An artificial neural network (ANN) is a collection of connected nodes, called neurons, structured in layers with connections between each of the neurons as seen in figure B.1. The general structure of ANNs is to have an input layer, hidden layers, and an output layer. Neurons in the input layer receive information/data to be processed and pass it on to the hidden layer. The information passes through the hidden layer and is processed. Finally, the output layer receives the processed data from the hidden layers and presents them as the output of the network. The hidden layers can vary in-depth, consisting of neurons connected from one layer to the next (or in some cases connected in the same layer). Each connection between neurons contains a weight which alters the value passed along it.

B.2 Feedforward Neural Network

The most common type of artificial neural network is a feedforward neural network (FNN). Information in a FNN is given to the input neurons and passed and processed forward through the network until it reaches the output layer. The manner in which a FNN processes information is by using activation functions and the weights on each of the connections. A neuron sums the product of all its input connections' values and their respective weights and uses the sum as input to the activation function. The activation function (which differs depending on the network) is used to introduce non-linearity

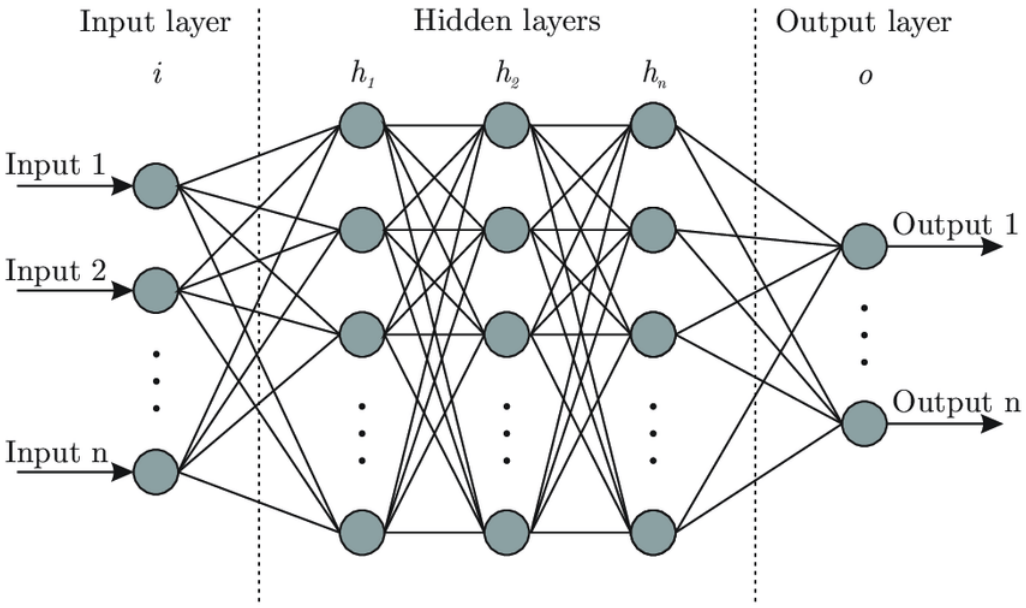


Figure B.1: Illustration of an artificial neural network with input, output, and hidden layers (Bre *et al.*, 2017). More specifically, this is an illustration of a feedforward neural network.

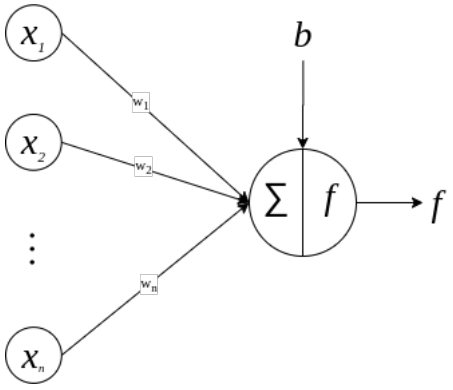


Figure B.2: Depiction of a simple neuron with inputs (x_1 to x_n) and their respective weights (w_1 to w_n). Some networks include a bias term for each layer, where the bias term is summed along with all input values.

into a network. It is also used to determine whether or not the neuron should fire or not (output a value or not). Some networks include a bias term for each layer, illustrated in figure B.2. The output of the activation function is then sent to the neurons connected to its output connections. Different activation functions can be used for different tasks, but some notable activation functions are ReLu (rectified linear unit), sigmoid, and hyperbolic tangent. In figure B.2 the output f of the neuron is calculated as follows:

$$f = (b + \sum_{i=1}^{n} x_i w_i) \tag{B.1}$$

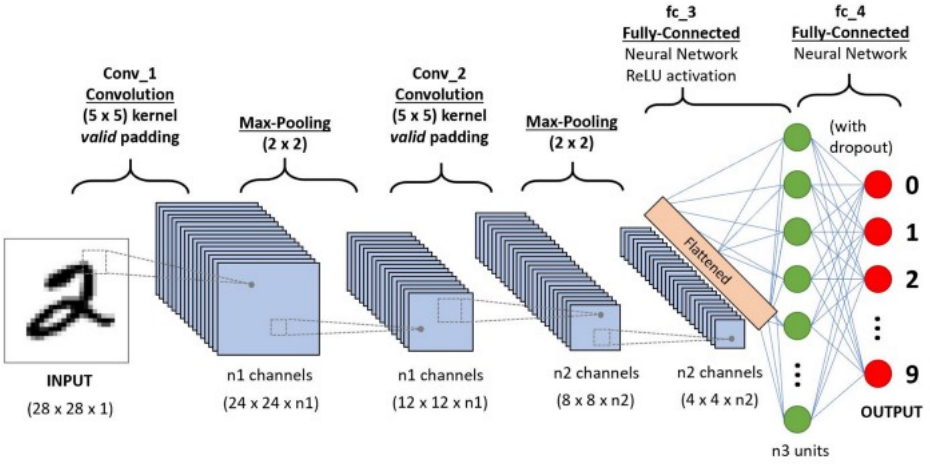


Figure B.3: A CNN architecture to classify handwritten digits (Saha, 2018)

Feedforward networks are used to approximate functions. The function approximation is made by training the network on known input, and output pairs denoted $(\overrightarrow{x}_i, \overrightarrow{y}_i)$. Input is given to the network, and an error is calculated by comparing the output to the desired output. This error is then used to change the weight values. This process is known as back-propagation and is done most commonly using gradient descent.

The goal of back-propagation is to find the weights and biases that would minimise the error. The process of adjusting the weights starts by calculating the error using a cost function, which takes as input the given output and expected output calculates the difference (most common is the mean squared distance between the two values). Each weight that contributed to the error is updated proportionally to how much it influenced the result. Updating weights is done iteratively, layer by layer, from the weights feeding into the output layer up to the weights coming from the input layer.

B.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are used to analyse images. Each layer in a CNN has a set of convolution kernels or filters that are used to calculate feature maps of the image. The feature map is calculated by sliding the kernel over the image and calculating correlation of the two matrices as seen in figure B.4. This process is referred to as cross-correlation. Each layer takes the feature map from the preceding layer and computes the cross-correlation with its kernel. This process continues until the desired dimension of feature map is acquired. Additional layers such as pooling or activation layers can be used in-between different convolution layers to make the network more robust and guard against overfitting the network. Pooling is the process of combining

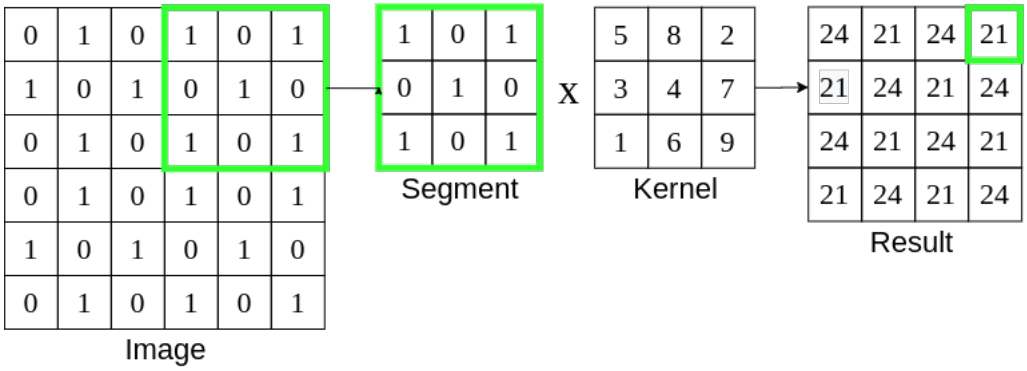


Figure B.4: Illustration of a typical CNN kernel being applied to an image

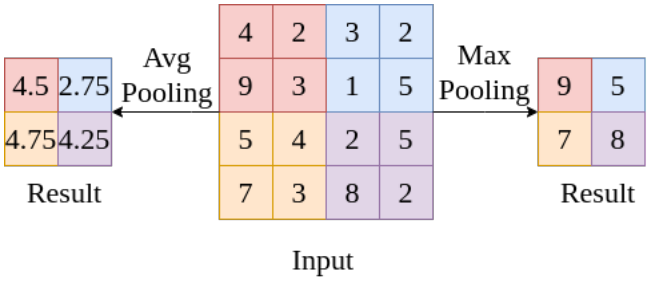


Figure B.5: Average and maximum pooling.

several pixels by either taking the average or the maximum and is visually shown in figure B.5. The activation function is performed on each value in the input image/feature map.

In essence, the kernels and layers extract features out of the input images, and the result is a feature vector. This vector is then used with a FNN to classify images in standard image classification scenarios. Training is implemented by adjusting each kernel's values according to an error calculated. Thus the kernel values can be seen as the CNN's weights.

Traditionally image processing was done with standard FNNs. The benefit of CNNs over FNNs is that the features that are extracted are shift and scale-invariant. The amount of training needed in order to get adequate results (which differ for different scenarios) is also much less for CNNs.

B.4 Recurrent Neural Networks

Recurrent neural networks work similar to the network it is derived from, feedforward network, with one major difference: it can look at prior inputs and learn from them. Figure B.6 gives a visual illustration, where x_n is an input, y_n an output, h_n a hidden state (which can be seen as its own FNN), and n the depth of the recurrent network. The input to RNNs is time-series data; data structured one after the other such as audio. The first piece of

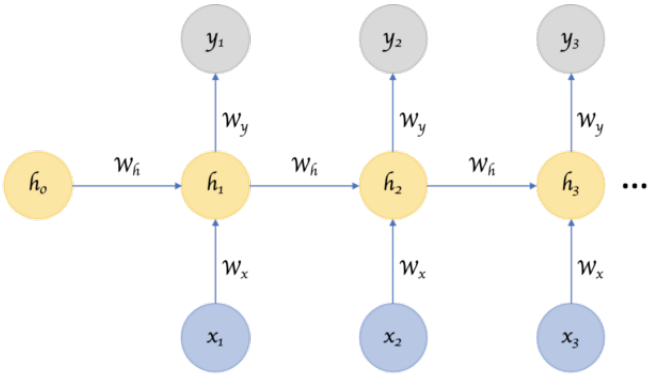


Figure B.6: A Recurrent Neural Network, with a hidden state that is meant to carry pertinent information from one input item in the series to others (Venkatachalam, 2019)

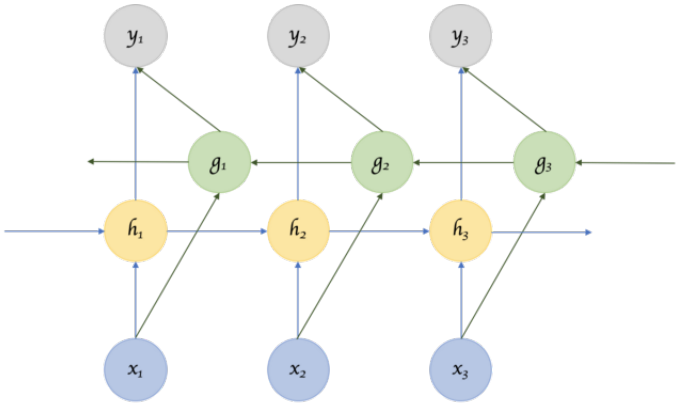


Figure B.7: A bi-directional recurrent neural network that allows for looking ahead, as well as at previous data to make a prediction (Venkatachalam, 2019)

the audio data would be fed into x_1, and an output will be derived at y_1 using the input x_1 and learned hidden weights from h_0 (which corresponds to a bias term) and h_1. The next time step, the data is given to x_1 is now passed onto x_2, and a new piece of data is presented at x_1. Again, the output from y_1 is calculated using x_1, h_0, and h_1. Similarly, y_2 is calculated using x_2, h_1, and h_2. However, the information acquired from h_1 corresponds to the input x_1, which in the time series data relates to the previous time step data. This process continues for any number of n and any length of input. By using this architecture, predictions can be made using previous information, which is important for audio processing. Training this network is like training n concurrent FNNs. There are many variations of RNNs, but particularly, bi-directional RNNs are of interest.

Similar to the standard RNN, bi-directional RNNs can look at both previous and future data in order to make a prediction, as seen in figure B.7. Being

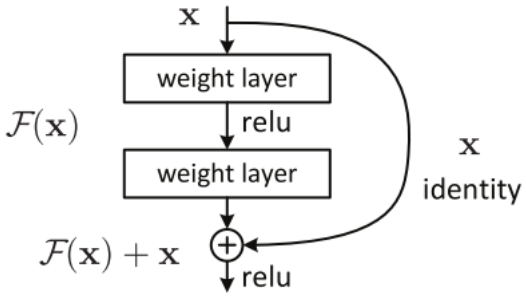


Figure B.8: A residual block.

able to look in both directions allows for better predictions with regards to speech analysis, as more information is present to predict what is said in the current time segment.

B.5 ResNet

The ResNet architecture first theorised by He *et al.* (2016*a*), is a convolutional neural network that makes use of something called residual blocks. Residual blocks, as seen in figure B.8, add skip-connections where the input to a network block (characterised by weights layer, a **relu** activation function, and another weights layer) is added to its output to form the new input for the next block. In figure B.9 the ResNet-34 network architecture is illustrated. Each **conv** block corresponds to a convolution block as described earlier with **A**x**A** defines a kernel of size **A** (meaning the kernel is an **A** by **A** matrix). The skip connections are indicated by arrows. In the case of the tablet application, a pre-trained ResNet-152 network was used. The 152 corresponds to the depths of the network, where the network in figure B.9 is only a 34 layer network, the ResNet-152 model used had 152 layers.

The deeper a network architecture is, the more time consuming it becomes to train it. The remarkable effect residual blocks have is that it reduces training time for deeper networks, increases accuracy, and reduces complexity.

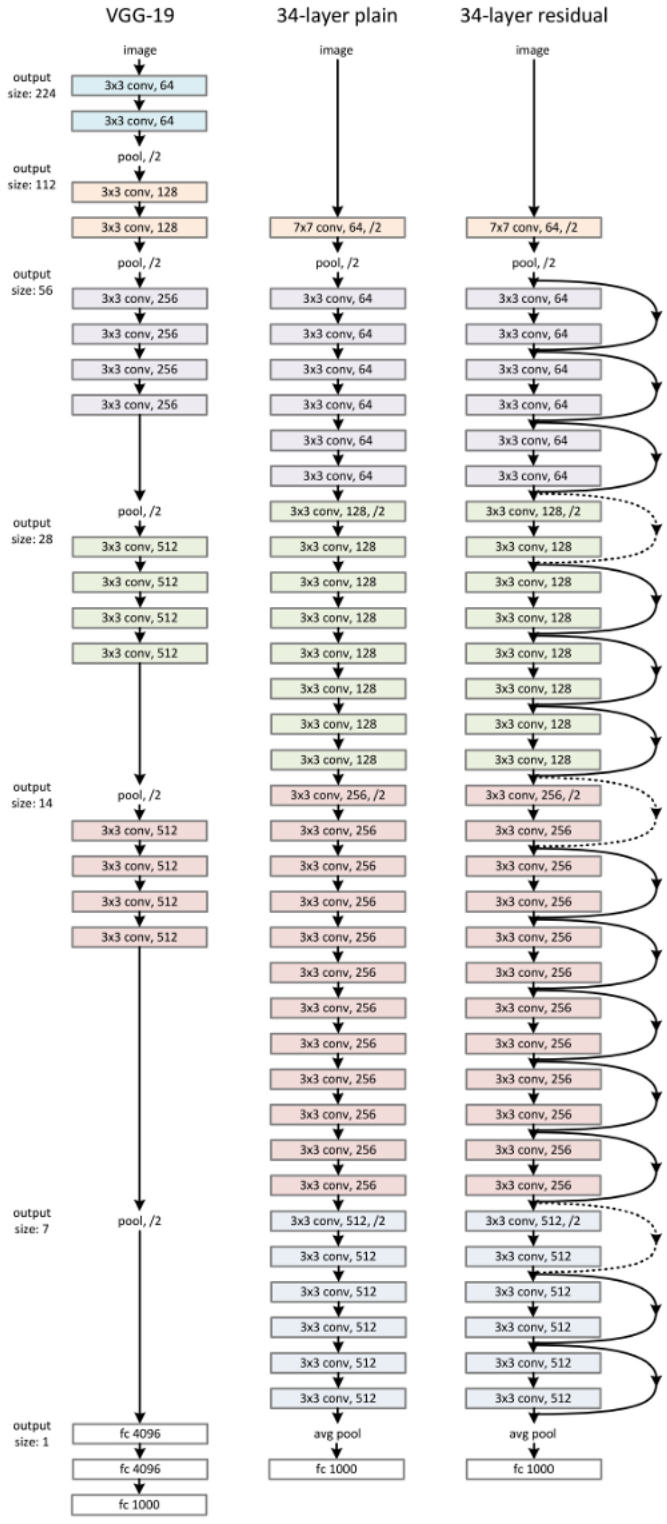


Figure B.9: ResNet-34 (right) compared to VGG19 network (left) and a normal deep CNN (middle) (He *et al.*, 2016*a*).

List of References

Agyei, S.B., van der Weel, F.R. and van der Meer, A.L. (2016 apr). Longitudinal study of preterm and full-term infants: High-density EEG analyses of cortical activity in response to visual motion. *Neuropsychologia*, vol. 84, pp. 89–104. ISSN 18733514.

Alloway, T.P. (2007 jan). Working memory, reading, and mathematical skills in children with developmental coordination disorder. *Journal of Experimental Child Psychology*, vol. 96, no. 1, pp. 20–36. ISSN 00220965.

Amod, Z., Cockcroft, K. and Soellaart, B. (2007 oct). Use of the 1996 Griffiths Mental Development Scales for infants: A pilot study with a Black, South African sample. *Journal of Child and Adolescent Mental Health*, vol. 19, no. 2, pp. 123–130. ISSN 17280591.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A.N., Han, T., Johannes, L.V., Jiang, B., Ju, C., Jun, B., Legresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J. and Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In: *33rd International Conference on Machine Learning, ICML 2016*, vol. 1, pp. 312–321. ISBN 9781510829008. `1512.02595`.

Anderson, P. (2002 jul). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology*, vol. 8, no. 2, pp. 71–82. ISSN 09297049.

Anwar, A. (2019). Difference between AlexNet, VGGNet, ResNet, and Inception | by Aqeel Anwar | Towards Data Science.
Available at: `https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96`

Aoki, S., Hashimoto, K., Mezawa, H., Hatakenaka, Y., Yasumitsu-Lovell, K., Suganuma, N., Ohya, Y., Wilson, P., Fernell, E., Kamio, Y. and Gillberg, C. (2018

jun). Development of a new screening tool for neuromotor development in children aged two, the neuromotor 5 min exam 2-year-old version (N5E2). *Brain and Development*, vol. 40, no. 6, pp. 445–451. ISSN 18727131.

Appalaraju, S. and Chaoji, V. (2017). Image similarity using Deep CNN and Curriculum Learning. `1709.08761`.

Avanzino, L., Pelosin, E., Vicario, C.M., Lagravinese, G., Abbruzzese, G. and Martino, D. (2016 dec). Time Processing and Motor Control in Movement Disorders. *Frontiers in Human Neuroscience*, vol. 10, p. 631. ISSN 1662-5161.

Baron, I.S., Anderson, P.J., Bauer, P.J., Leventon, J.S., Varga, N.L., Baron, I.S. and Anderson, P.J. (2012 dec). Neuropsychological Assessment of Preschoolers.

Baron, I.S. and Leonberger, K.A. (2012 dec). Assessment of intelligence in the preschool period.

Behne, T., Liszkowski, U., Carpenter, M. and Tomasello, M. (2012 sep). Twelve-month-olds' comprehension and production of pointing. *British Journal of Developmental Psychology*, vol. 30, no. 3, pp. 359–375. ISSN 0261510X.

Bhavnani, S., Mukherjee, D., Dasgupta, J., Verma, D., Parameshwaran, D., Divan, G., Sharma, K.K., Thiagarajan, T. and Patel, V. (2019). Development, feasibility and acceptability of a gamified cognitive DEvelopmental assessment on an E-Platform (DEEP) in rural Indian pre-schoolers: a pilot study. *Global Health Action*, vol. 12, no. 1. ISSN 16549880.

Bishop, D.V. and Adams, C. (1990). A Prospective Study of the Relationship between Specific Language Impairment, Phonological Disorders and Reading Retardation. *Journal of Child Psychology and Psychiatry*, vol. 31, no. 7, pp. 1027–1050. ISSN 14697610.

Bishop, D.V. and Edmundson, A. (1987). Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders*, vol. 52, no. 2, pp. 156–173. ISSN 00224677.

Bre, F., Gimenez, J. and Fachinotti, V. (2017 11). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, vol. 158.

Brown, T.T. and Jernigan, T.L. (2012). Brain development during the preschool years.

Bruininks, R.H. and Bruininks, B.D. (2005). BOT-2 Bruininks-Oseretsky Test of Motor Proficiency Ed. 2.

Bruininks, R.H. *et al.* (1978). *Bruininks-Oseretsky test of motor proficiency*. American Guidance Service Circle Pines, MN.

Cairney, J., Veldhuizen, S. and Szatmari, P. (2010 jul). Motor coordination and emotional-behavioral problems in children.

Cameron, C.E., Brock, L.L., Murrah, W.M., Bell, L.H., Worzalla, S.L., Grissmer, D. and Morrison, F.J. (2012). Fine Motor Skills and Executive Function Both Contribute to Kindergarten Achievement. *Child Development*, vol. 83, no. 4, pp. 1229–1244. ISSN 00093920.

Carreiras, M., Quiñones, I., Hernández-Cabrera, J.A. and Duñabeitia, J.A. (2015). Orthographic coding: Brain activation for letters, symbols, and digits. *Cerebral Cortex*, vol. 25, no. 12, pp. 4748–4760. ISSN 14602199.

Casey, B.J., Tottenham, N., Liston, C. and Durston, S. (2005). Imaging the developing brain: What have we learned about cognitive development? In: *Trends in Cognitive Sciences*, vol. 9, pp. 104–110. Elsevier Ltd. ISSN 13646613.

Chiong, C. and Shuler, C. (2010). The Joan Ganz Cooney Center at Sesame Workshop. Learning: Is there an app for that?

Chugani, H.T. (1998). A critical period of brain development: Studies of cerebral glucose utilization with PET. In: *Preventive Medicine*, vol. 27, pp. 184–188. Academic Press Inc. ISSN 00917435.

Claus, F., Rosales, H.G., Petrick, R., Hain, H.-u. and Hoffman, R. (2013). A Survey about ASR for Children. In: *Proceedings of SLaTE*, pp. 26–30.

Cohen, E.J., Bravi, R., Bagni, M.A. and Minciacchi, D. (2018 oct). Precision in drawing and tracing tasks: Different measures for different aspects of fine motor control. *Human Movement Science*, vol. 61, pp. 177–188. ISSN 18727646.

Conti-Ramsden, G. and Durkin, K. (2007 feb). Phonological short-term memory, language and literacy: Developmental relationships in early adolescence in young people with SLI. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 48, no. 2, pp. 147–156. ISSN 00219630.

Conti-Ramsden, G. and Durkin, K. (2012 dec). Language development and assessment in the preschool period.

Conti-Ramsden, G., Durkin, K., Simkin, Z. and Knox, E. (2009 jan). Specific language impairment and school outcomes. I: Identifying and explaining variability at the end of compulsory education. *International Journal of Language and Communication Disorders*, vol. 44, no. 1, pp. 15–35. ISSN 13682822.

Crutchley, A., Botting, N. and Conti-Ramsden, G. (1997). Bilingualism and specific language impairment in children attending language units. *International Journal of Language and Communication Disorders*, vol. 32, no. 2, pp. 267–276. ISSN 13682822.

De Beer, R. (2019). Research meeting with speech therapist.

Dewey, D., Cantell, M. and Crawford, S.G. (2007 mar). Motor and gestural performance in children with autism spectrum disorders, developmental coordination disorder, and/or attention deficit hyperactivity disorder. *Journal of the International Neuropsychological Society*, vol. 13, no. 2, pp. 246–256. ISSN 14697661.

Duncan, G.J. and Brooks-Gunn, J. (2000 jan). Family poverty, welfare reform, and child development. *Child Development*, vol. 71, no. 1, pp. 188–196. ISSN 00093920.

Dunn, L.M. and Dunn, D.M. (2007). PPVT-4 Peabody Picture Vocabulary Test 4th Edition.

Dunn, L.M. and Dunn, D.M. (2009). British Picture Vocabulary Scale: 3rd Edition - BPVS III.

Durkin, K. and Conti-Ramsden, G. (2007 sep). Language, social behavior, and the quality of friendships in adolescents with and without a history of specific language impairment. *Child Development*, vol. 78, no. 5, pp. 1441–1457. ISSN 00093920.

Engle, P.L. and Black, M.M. (2008 jun). The effect of poverty on child development and educational outcomes.

Erez, O., Gordon, C.R., Sever, J., Sadeh, A. and Mintz, M. (2004 jan). Balance dysfunction in childhood anxiety: Findings and theoretical approach. *Journal of Anxiety Disorders*, vol. 18, no. 3, pp. 341–356. ISSN 08876185.

Evans, G.W. (2006 jan). Child development and the physical environment.

Falter, C.M. and Noreika, V. (2011). Interval Timing Deficits and Abnormal Cognitive Development. *Frontiers in Integrative Neuroscience*, vol. 5, no. June, pp. 1–2. ISSN 1662-5145.

Fawcett, A.J. and Nicolson, R.I. (1995 sep). Persistent deficits in motor skill of children with dyslexia. *Journal of Motor Behavior*, vol. 27, no. 3, pp. 235–240. ISSN 19401027.

Fayek, H.M. (2016). Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between.

Folio, M.R. and Fewell, R.R. (2000). PDMS-2 Peabody Developmental Motor Scales 2nd Edition.

Francis-Lyon, P., Attiga, Y., Manjunath, R., Ramasubramanian, U., Chaudhuri, V., Nguyen, T., Xu, X., Zeng, S., Abubakar, A. and Newton, C.R. (2017 dec). Tablet app for child cognitive assessment in low and middle income countries. In: *GHTC 2017 - IEEE Global Humanitarian Technology Conference, Proceedings*, vol. 2017-Janua, pp. 1–5. Institute of Electrical and Electronics Engineers Inc. ISBN 9781509060467.

Frankenburg, W.K., Dodds, J., Archer, P., Shapiro, H. and Bresnick, B. (1992). The Denver II: A major revision and restandardization of the Denver Developmental Screening Test. *Pediatrics*, vol. 89, no. 1, pp. 91–97. ISSN 00314005.

Geist, E.a. (2012). A qualitative examination of two year-olds interaction with tablet based interactive technology. *Journal of Instructional Psychology*, vol. 39, no. 1, pp. 26–35. ISSN 0094-1956.

Gerosa, M., Giuliani, D., Narayanan, S. and Potamianos, A. (2009). A review of ASR technologies for children's speech. In: *Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09*, pp. 1–8. ACM Press, New York, New York, USA. ISBN 9781605586908.

Goldfield, B.A. and Reznick, J.S. (1990 feb). Early Lexical Acquisition: Rate, Content, And The Vocabulary Spurt. *Journal of Child Language*, vol. 17, no. 1, pp. 171–183. ISSN 14697602.

Gorey, K.M. (2001 mar). Early Childhood Education: A Meta-Analytic Affirmation of the Short- and Long-Term Benefits of Educational Opportunity. *School Psychology Quarterly*, vol. 16, no. 1, pp. 9–30. ISSN 10453830.

Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: *ACM International Conference Proceeding Series*, vol. 148, pp. 369–376. ISBN 1595933832.

Gray, S.S., Willett, D., Lu, J., Pinto, J., Maergner, P. and Bodenstab, N. (2014). Child Automatic Speech Recognition for US English : Child Interaction with living-room-electronic-devices. *Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI 2014)*, , no. Wocci, pp. 21–26.

Hannun, A. (2017 nov). Sequence Modeling with CTC. *Distill*, vol. 2, no. 11, p. e8. ISSN 2476-0757.
Available at: `https://distill.pub/2017/ctc`

Harvey, P.D. (2019). Domains of cognition and their assessment. *Dialogues in Clinical Neuroscience*, vol. 21, no. 3, pp. 227–237. ISSN 12948322.

He, K., Zhang, X., Ren, S. and Sun, J. (2016*a*). Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778. ISBN 9781467388504. ISSN 10636919. `1512.03385`.

He, K., Zhang, X., Ren, S. and Sun, J. (2016 dec*b*). Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778. IEEE Computer Society. ISBN 9781467388504. ISSN 10636919. `1512.03385`.

Henderson, S.E., Sugden, D.A. and Barnett, A.L. (2007). Movement Assessment Battery for Children - Second Edition (Movement ABC-2) | Pearson Assessment.

Hoff, E. (2009). *Language Development (4th edition)*.

Howard, S.J. and Melhuish, E. (2017 jun). An Early Years Toolbox for Assessing Early Executive Function, Language, Self-Regulation, and Social Development: Validity, Reliability, and Preliminary Norms. *Journal of Psychoeducational Assessment*, vol. 35, no. 3, pp. 255–275. ISSN 07342829.

Howard, S.J. and Okely, A.D. (2015 sep). Catching Fish and Avoiding Sharks: Investigating Factors That Influence Developmentally Appropriate Measurement of Preschoolers' Inhibitory Control. *Journal of Psychoeducational Assessment*, vol. 33, no. 6, pp. 585–596. ISSN 07342829.

Jin, X., Sun, Y., Jiang, F., Ma, J., Morgan, C. and Shen, X. (2007 jun). "Care for development" intervention in rural China: A prospective follow-up study. *Journal of Developmental and Behavioral Pediatrics*, vol. 28, no. 3, pp. 213–218. ISSN 0196206X.

Kent, R.D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: evidence from acoustic studies. *Journal of Speech and Hearing Research*, vol. 19, no. 3, pp. 421–445. ISSN 00224685.

Kucirkova, N. (2014). iPads in early education: Separating assumptions and evidence.

Kuperman, V., Stadthagen-Gonzalez, H. and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, vol. 44, no. 4, pp. 978–990. ISSN 15543528.

Kurdek, L.A. and Sinclair, R.J. (2001). Predicting reading and mathematics achievement in fourth-grade children from kindergarten readiness scores. *Journal of Educational Psychology*, vol. 93, no. 3, pp. 451–455. ISSN 00220663.

Laughton, B., Springer, P.E., Grove, D., Seedat, S., Cornell, M., Kidd, M., Madhi, S.A. and Cotton, M.F. (2010). Longitudinal developmental profile of children from low socio-economic circumstances in Cape Town, using the 1996 Griffiths Mental Development Scales | Laughton | South African Journal of Child Health.

Law, J., Boyle, J., Harris, F., Harkness, A. and Nye, C. (1998). Screening for speech and language delay: A systematic review of the literature.

Leonard, L.B., Weismer, S.E., Miller, C.A., Francis, D.J., Bruce Tomblin, J. and Kail, R.V. (2007 apr). Speed of processing, working memory, and language impairment in children. *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 2, pp. 408–428. ISSN 10924388.

Light, J.C., Roberts, B., Dimarco, R. and Greiner, N. (1998 mar). Augmentative and alternative communication to support receptive and expressive communication for people with autism. *Journal of Communication Disorders*, vol. 31, no. 2, pp. 153–180. ISSN 00219924.

Lowe, D.G. (1999). Object recognition from local scale-invariant features. In: *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157.

Martin, N.A. and Brownell, R. (2010*a*). EOWPVT-4: Expressive One-Word Picture Vocabulary Test - Fourth Edition.

Martin, N.A. and Brownell, R. (2010*b*). ROWPVT-4: Receptive One-Word Picture Vocabulary TestFourth Edition.

McCarron, L.T. (1997). MAND : McCarron assessment of neuromuscular development, fine and gross motor abilities.

McMurray, B. and Aslin, R.N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, vol. 95, no. 2. ISSN 00100277.

Mukherjee, D., Bhavnani, S., Swaminathan, A., Verma, D., Parameshwaran, D., Divan, G., Dasgupta, J., Sharma, K., Thiagarajan, T.C. and Patel, V. (2020). Proof of Concept of a Gamified DEvelopmental Assessment on an E-Platform (DEEP) Tool to Measure Cognitive Development in Rural Indian Preschool Children. *Frontiers in Psychology*, vol. 11. ISSN 16641078.

Murray, G.K., Veijola, J., Moilanen, K., Miettunen, J., Glahn, D.C., Cannon, T.D., Jones, P.B. and Isohanni, M. (2006 jan). Infant motor development is associated with adult cognitive categorisation in a longitudinal birth cohort study. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 47, no. 1, pp. 25–29. ISSN 00219630.

Nacher, V., Jaen, J., Navarro, E., Catala, A. and González, P. (2015 jan). Multitouch gestures for pre-kindergarten children. *International Journal of Human Computer Studies*, vol. 73, pp. 37–51. ISSN 10959300.

Nampijja, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A.M. and Alcock, K.J. (2010 mar). Adaptation of western measures of cognition for assessing 5-year-old semi-urban Ugandan children. *British Journal of Educational Psychology*, vol. 80, no. 1, pp. 15–30. ISSN 00070998.

Newton, T.J. and Joyce, A.P. (2012). *Human Perspectives*.
Available at: `https://books.google.co.za/books/about/Human_Perspectives.html?id=bWrESgAACAAJ&redir_esc=y`

Noreika, V., Falter, C.M. and Rubia, K. (2013). Timing deficits in attention-deficit/hyperactivity disorder (ADHD): Evidence from neurocognitive and neuroimaging studies. *Neuropsychologia*, vol. 51, no. 2, pp. 235–266. ISSN 00283932.

Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210.

Papanicolaou, A.C., Simos, P.G., Breier, J.I., Zouridakis, G., Willmore, L.J., Wheless, J.W., Constantinou, J.E., Maggio, W.W. and Gormley, W.B. (1999 jan). Magnetoencephalographic mapping of the language-specific cortex. *Journal of Neurosurgery*, vol. 90, no. 1, pp. 85–93. ISSN 00223085.

Paul, R. (1996 may). Clinical Implications of the Natural History of Slow Expressive Language Development. *American Journal of Speech-Language Pathology*, vol. 5, no. 2, pp. 5–21. ISSN 1058-0360.

Piek, J.P., Barrett, N.C., Smith, L.M., Rigoli, D. and Gasson, N. (2010 oct). Do motor skills in infancy and early childhood predict anxious and depressive symptomatology at school age? *Human Movement Science*, vol. 29, no. 5, pp. 777–786. ISSN 01679457.

Piek, J.P., Hands, B. and Licari, M.K. (2012 dec). Assessment of motor functioning in the preschool period.

Pitcher, T.M., Piek, J.P. and Hay, D.A. (2003 feb). Fine and gross motor ability in males with ADHD. *Developmental Medicine and Child Neurology*, vol. 45, no. 8, pp. 525–535. ISSN 00121622.

Pitchford, N.J. and Outhwaite, L.A. (2016 oct). Can touch screen tablets be used to assess cognitive and motor skills in early years primary school children? A cross-cultural study. *Frontiers in Psychology*, vol. 7, no. OCT. ISSN 16641078.

Rey Otero, I. and Delbracio, M. (2014). Anatomy of the SIFT Method. *Image Processing On Line*, vol. 4, pp. 370–396. ISSN 2105-1232.

Rhode, A. (2019). Research meeting with neurodevelopmentally trained occupational therapist.

Richmond, J.B., Health, C., Cameron, J. and Ph, D. (2016). The Timing and Quality of Early Experiences Combine to Shape Brain Architecture. *The American Economist*, vol. 6, no. 1, pp. 29–29. ISSN 0569-4345.

Ruckligde, W. (1996). Efficient visual recognition using the Hausdorff distance. In: Rucklidge, W. (ed.), *Lecture Notes in Computer Science*, vol. 1173 of *Lecture Notes in Computer Science*, pp. 27–42. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-61993-2.

Saaristo-Helin, K., Kunnari, S. and Savinainen-Makkonen, T. (2011 aug). Phonological development in children learning Finnish: A review. *First Language*, vol. 31, no. 3, pp. 342–363. ISSN 01427237.

Sabanathan, S., Wills, B. and Gladstone, M. (2015). Child development assessment tools in low-income and middle-income countries: How can we use them more appropriately? *Archives of Disease in Childhood*, vol. 100, no. 5, pp. 482–488. ISSN 14682044.

Sachdev, P.S., Blacker, D., Blazer, D.G., Ganguli, M., Jeste, D.V., Paulsen, J.S. and Petersen, R.C. (2014). Classifying neurocognitive disorders: The DSM-5 approach. *Nature Reviews Neurology*, vol. 10, no. 11, pp. 634–642. ISSN 17594766.

Saha, S. (2018 Dec). *A CNN sequence to classify handwritten digits*. Medium. Available at: `https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53`

Schaefer, B., Bowyer-Crane, C., Herrmann, F. and Fricke, S. (2016 jun). Development of a tablet application for the screening of receptive vocabulary skills in multilingual children: A pilot study. *Child Language Teaching and Therapy*, vol. 32, no. 2, pp. 179–191. ISSN 14770865.

Schipke, C.S. and Kauschke, C. (2011 feb). Early word formation in German language acquisition: A study on word formation growth during the second and third years. *First Language*, vol. 31, no. 1, pp. 67–82. ISSN 01427237.

Schuster, M. and Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681. ISSN 1053587X.

Seeff-Gabriel, B., Chiat, S. and Roy, P. (2008). Early Repetition Battery (ERB) | Pearson Assessment.

Seljan, S. and Dunder, I. (2014). Combined automatic speech recognition and machine translation in business correspondence domain for english-croatian. *International Journal of Industrial and Systems Engineering*, vol. 8, no. 11, pp. 1980 – 1986. ISSN eISSN: 1307-6892.
Available at: `https://publications.waset.org/vol/95`

Semel, E., Wiig, E.H. and Secord, W. (2006). Clinical evaluation of language fundamentals - preschool 2 UK edition.

Slater, L.M., Hillier, S.L. and Civetta, L.R. (2010). The Clinimetric Properties of Performance-Based Gross Motor Tests Used for Children with Developmental Coordination Disorder: A Systematic Review.

Smyth, M.M. and Anderson, H.I. (2000 sep). Coping with clumsiness in the school playground: Social and physical play in children with coordination impairments. *British Journal of Developmental Psychology*, vol. 18, no. 3, pp. 389–413. ISSN 0261510X.

St. Clair, M.C., Durkin, K., Conti-Ramsden, G. and Pickles, A. (2010 mar). Growth of reading skills in children with a history of specific language impairment: The role of autistic symptomatology and language-related abilities. *British Journal of Developmental Psychology*, vol. 28, no. 1, pp. 109–131. ISSN 0261510X.

St Clair, M.C., Pickles, A., Durkin, K. and Conti-Ramsden, G. (2011 mar). A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment (SLI). *Journal of Communication Disorders*, vol. 44, no. 2, pp. 186–199. ISSN 00219924.

Steven Barnett, W. (1998). Long-term cognitive and academic effects of early childhood education on children in poverty. In: *Preventive Medicine*, vol. 27, pp. 204–207. Academic Press Inc. ISSN 00917435.

Sudfeld, C.R., McCoy, D.C., Fink, G., Muhihi, A., Bellinger, D.C., Masanja, H., Smith, E.R., Danaei, G., Ezzati, M. and Fawzi, W.W. (2015 dec). Malnutrition and its determinants are associated with suboptimal cognitive, communication, and motor development in Tanzanian children. *Journal of Nutrition*, vol. 145, no. 12, pp. 2705–2714. ISSN 15416100.

Thal, D.J. and Katich, J. (1996). Predicaments in early identification of specific language impairment: Does the early bird always catch the worm? In: *Assessment of communication and language*, pp. 1–28.

Tincoff, R. and Jusczyk, P.W. (1999 mar). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, vol. 10, no. 2, pp. 172–175. ISSN 09567976.

Tyack, D. and Ingram, D. (1977 jun). Children's production and comprehension of questions. *Journal of Child Language*, vol. 4, no. 2, pp. 211–224. ISSN 14697602.

Van Der Walt, A. (2019). Research meeting with neurodevelopmental paediatrician.

Van Gemmert, A.W. and Teulings, H.L. (2006 oct). Advances in graphonomics: Studies on fine motor control, its development and disorders. *Human Movement Science*, vol. 25, no. 4-5, pp. 447–453. ISSN 01679457.

Venkatachalam, M. (2019 Mar). *A Recurrent Neural Network, with a hidden state that is meant to carry pertinent information from one input item in the series to others.* Medium.
Available at: `https://towardsdatascience.com/recurrent-neural-networks-d4642c9bc7ce`

Viding, E., Spinath, F.M., Price, T.S., Bishop, D.V., Dale, P.S. and Plomin, R. (2004 feb). Genetic and environmental influence on language impairment in 4-year-old same-sex and opposite-sex twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 45, no. 2, pp. 315–325. ISSN 00219630.

Vimercati, S.L., Galli, M., Stella, G., Caiazzo, G., Ancillao, A. and Albertini, G. (2015 mar). Clumsiness in fine motor tasks: Evidence from the quantitative drawing evaluation of children with Down Syndrome. *Journal of Intellectual Disability Research*, vol. 59, no. 3, pp. 248–256. ISSN 13652788.

Wehrmann, S., Chiu, T., Reid, D. and Sinclair, G. (2006). Evaluation of occupational therapy school-based consultation service for students with fine motor difficulties.

WHO (2018). Improving early childhood development: WHO guideline FREQUENTLY ASKED QUESTIONS. Tech. Rep..
Available at: `www.who.int/maternal_child_adolescent/child/summary_guideline_`

Williams, K.T. (2007). EVT-2 Expressive Vocabulary Test Second Edition.